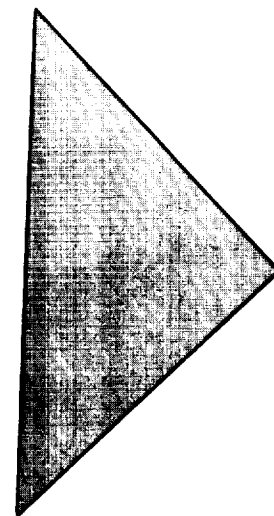
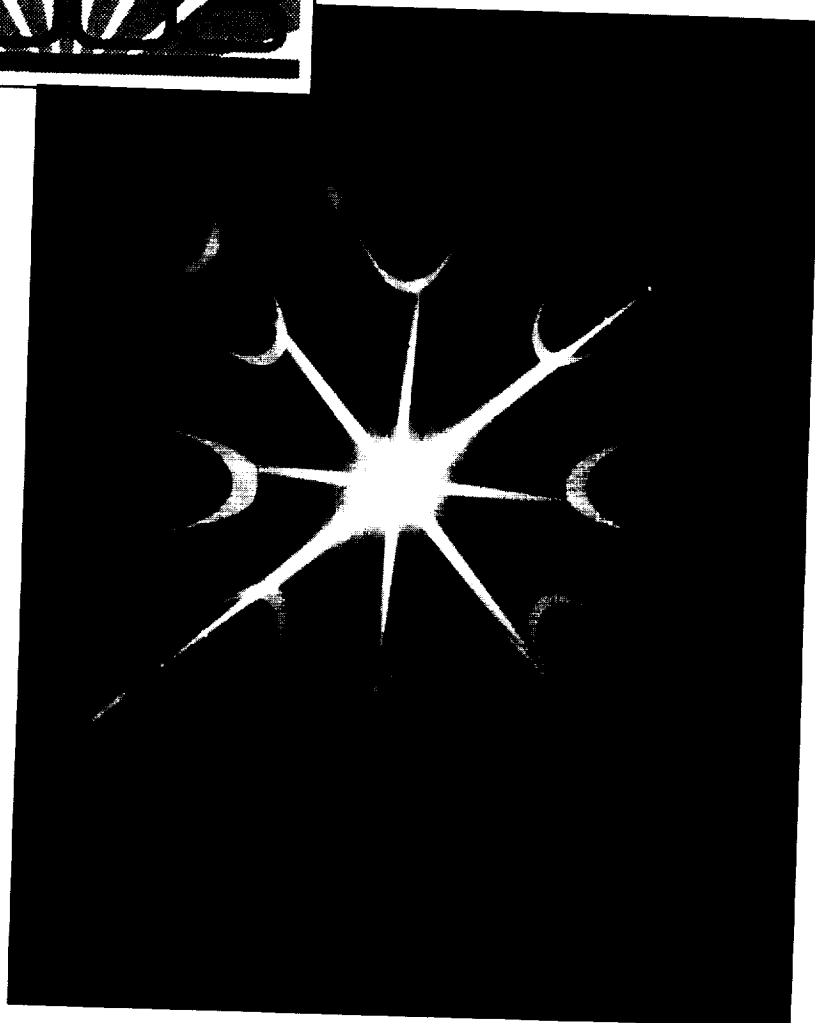
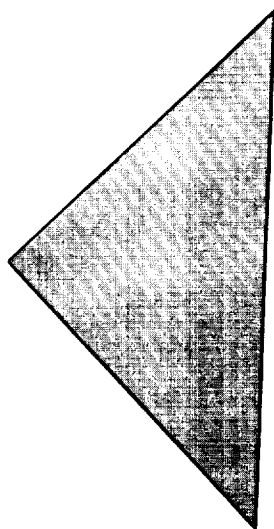
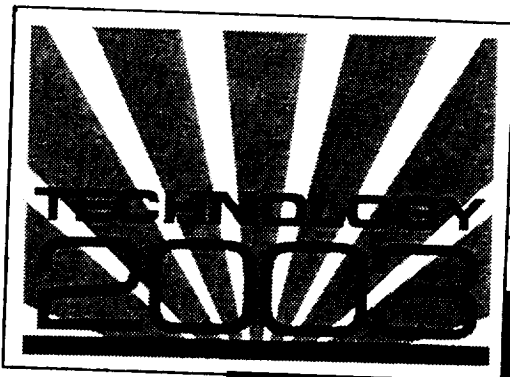


*NASA Conference Publication 3249 - Vol-2
Volume Two*



Conference Proceedings

The Fourth National Technology Transfer Conference & Exposition
December 7-9, 1993 • Anaheim, CA

*Sponsored by NASA, NASA Tech Briefs Magazine
and the Technology Utilization Foundation*



ACKNOWLEDGMENTS

The Technology 2003 Conference Management would like to thank the following individuals who generously contributed their time to serve as session moderators:

Dr. Josephine Covino
Head, Applied Mechanics Branch
Naval Air Warfare Center
Weapons Division
China Lake, CA

Dr. Hamed M. El-Bisi
Chief Scientist
Army Research Laboratory
Materials Directorate
Watertown, MA

Donald G. Foster
Head, Mechanical Processes Dept.
Lawrence Berkeley Laboratory
Berkeley, CA

Linda Geissinger
Program Analyst
Environmental Management
McClellan Air Force Base
McClellan AFB, CA

Arif Husain
Technology Transfer Office
NASA Resident Office
Jet Propulsion Laboratory
Pasadena, CA

Anton L. Interbitzen
Deputy Assistant Director
for Research
U.S. Geological Survey
Reston, VA

Diana C. Jackson
Technology Transfer Representative
Naval Command, Control and Ocean
Surveillance Center
San Diego, CA

Geoffrey S. Lee
Technology Transfer Officer
NASA Ames Research Center
Office of Commercial and
Community Programs
Moffett Field, CA

Michael L. Mastracci
Senior Environmental Engineer
U.S. Environmental Protection
Agency
Washington, DC

H. Dana Moran
Manager, Research and Technology
Applications
National Renewable Energy
Laboratory
Golden, CO

Dr. Mike Sullivan
Technology Transfer Manager
Naval Air Warfare Center
Point Mugu, CA

Dr. R. Michael Templeton
President
Templeton and Associates, Inc.
San Diego, CA

Diana L. West
Associate Program Leader
Technology Transfer Initiatives
Program
Lawrence Livermore National
Laboratory
Livermore, CA

TABLE OF CONTENTS

Symposia:

Artificial Intelligence	1
Refining Fuzzy Logic Controllers with Machine Learning	3
A Fuzzy Classifier System for Process Control	7
Fuzzy-Neural Control of an Aircraft Tracking Camera Platform	17
Empirical Modeling for Intelligent, Real-Time Manufacture Control	24
Neural Network Wavelet Technology: A Frontier of Automation	34
New Approaches for Real-Time Decision Support Systems	40
Knowledge-Based Commodity Distribution Planning	46
A Hypertext System That Learns From User Feedback	55
Computer-Aided Design and Engineering	63
Common Modeling System for Digital Simulation	65
Analytical Design Package - ADP2: A Computer-Aided Engineering Tool for Aircraft Transparency Design	80
Assembly Flow Simulation of a Radar	93
CONFIG: Integrated Engineering of Systems and Their Operation	97
Computer Hardware	105
Spacecraft Onboard Information Extraction Computer (SOBIEC)	107
Pen-Based Computers: Computers Without Keys	117
"The Vertical"	124
A Systems Approach to Computer-Based Training	130
Computer Software	137
Automatic Translation Among Spoken Languages	139
A PC Program to Optimize System Configuration for Desired Reliability at Minimum Cost	143
Evolving Software Re-engineering Technology for the Emerging Innovative-Competitive Era	151
A High-Speed Linear Algebra Library With Automatic Parallelism	161
Information Management	169
High-Speed Data Search	171
Database Tomography for Commercial Application	181
Automated Mainframe Data Collection in a Network Environment	190
Beginning the 21st Century With Advanced Automatic Parts Identification (API)	198
Photonics	205
Optical Processing for Semiconductor Device Fabrication	207
A Scanning Defect Mapping System for Semiconductor Characterization	216
Neutral Ion Sources in Precision Manufacturing	223
High-Power Diode Lasers for Solid-State Laser Pumps	233
Flexible Manufacturing for Photonics Device Assembly	246

TABLE OF CONTENTS

Robotics	253
Application of Dexterous Space Robotics Technology to Myoelectric Prostheses	255
U.S. Navy Omni-Directional Vehicle (ODV) Development Program	269
Applying Robotics to HAZMAT	279
Advanced Teleoperation: Technology Innovations and Applications	288
Test and Measurement	293
Continuous Measurement of Aircraft Wing Icing	295
Electromagnetic Probe Technique for Fluid Flow Measurements	301
Computerized Ultrasonic Testing System (CUTS) for In-Process Thickness Determination ..	311
Microwave Sensor for Ice Detection	318
A Versatile Nondestructive Evaluation Imaging Workstation	326
A New High-Speed IR Camera System	332
Universal Signal Conditioning Amplifier (USCA)	342
The Constant Current Loop: A New Paradigm for Resistance Signal Conditioning	349
Video and Imaging Technology	363
A Multimedia Adult Literacy Package Combining NASA Technology, Recursive ID Theory and Authentic Instruction Theory	365
Mapping, Analysis and Planning System for the Kennedy Space Center	375
Remote Sensing for Hurricane Andrew Impact Assessment	388
Remote Sensing for Urban Planning	389
Remote Sensing and the Mississippi High Accuracy Reference Network	394
A Visual Detection Model for DCT Coefficient Quantization	404
Voice and Video Transmission Using XTP and FDDI	416
An Intelligent Interactive Visual Database Management System for Space Shuttle Closeout Image Management	422
The Trustworthy Digital Camera: Restoring Credibility to the Photographic Image	430
Virtual Reality/Simulation	437
Virtual Reality in Medical Education and Assessment	439
Technology Transfer of Operator-In-The-Loop Simulation	444
High-Performance Real-Time Flight Simulation at NASA Langley	456
The Effects of Above Real-Time Training (ARTT) in an F-16 Simulator	463

For information regarding additional copies of the
Technology 2003 Conference Proceedings, please contact:

The Technology Utilization Foundation
41 East 42nd Street, Suite 921
New York, NY 10017
ph.: 212/490-3999; fax: 212/986-7864

PRECEDING PAGE BLANK NOT FILMED

1915

N94- 32421

REFINING FUZZY LOGIC CONTROLLERS WITH MACHINE LEARNING

Hamid R. Berenji
Intelligent Inference Systems Corp.
AI Research Branch, MS: 269-2
NASA Ames Research Center
Mountain View, CA 94035
berenji@ptolemy.arc.nasa.gov

2-03
2484
P-4

ABSTRACT

In this paper, we describe GARIC (Generalized Approximate Reasoning-Based Intelligent Control) architecture which learns from past performance and modifies the labels in the fuzzy rules to improve performance. It uses fuzzy reinforcement learning which is a hybrid method of fuzzy logic and reinforcement learning. This technology can simplify and automate the application of fuzzy logic control to a variety of systems. GARIC has been applied in simulation studies of the Space Shuttle rendezvous and docking experiments. It has the potential of being applied in other aerospace systems as well as in consumer products such as appliances, cameras, and cars.

INTRODUCTION

Future generation of intelligent systems are expected to demonstrate a high degree of autonomy in their operations. For example, the future controllers for the Space Shuttle in-orbit operations should consume less fuel, be capable of performing more difficult maneuvers and rendezvous missions, and eliminate jet over-firings which can result in payload contamination. Fuzzy logic control can play an important role in development of intelligent systems for space applications [4], [6], [7], [8] where the experience of human experts can be modeled and used.

Fuzzy logic controllers generally use rules containing fuzzy terms such as small, medium, and large which can be mathematically represented using membership functions. The membership values range between zero (for non-membership) to one (for full membership). For example, a temperature reading of 85 degrees may be given a membership value of .9 in a fuzzy set *hot*.

A difficulty in design of these systems relates to fine-tuning the membership functions of the labels used in the rules. A few approaches have been recently suggested which use neural networks to define and fine-tune the membership functions (e.g., [5]). However, these have been mostly off-line and supervised learning approaches. In [1], [2], [3] the idea of using reinforcement learning for developing fuzzy membership functions has been proposed and two architectures, ARIC and GARIC, have been developed. After successful applications of these architectures to cart-pole balancing and truck backing, the performances of these algorithms in the attitude control and rendezvous docking missions of the Space Shuttle are being studied [4]. In this paper, we discuss some of the lessons learned in applying the GARIC architecture to a complex system such as the simulation of in-orbit operation of the Space Shuttle.

THE GARIC ARCHITECTURE

In some sense, GARIC emulates the way that humans learn to become experts in performing a task. For example, in learning to play tennis, a novice player first learns a number of general rules for playing this game. These rules may include how to hold the racket, how to move from a place to the next depending on the location of the opponent and the direction of the ball movement, etc. In GARIC, such a process helps in the definition of fuzzy control rules. After these general rules, which are approximate by their nature, have been learned by the novice tennis player, then he or she starts to practice. It can be argued that by practicing more, the player sharpens his or her skills in order to produce higher reinforcements (i.e., to win more games). This process in GARIC refers to refining the fuzzy control rules and in particular, changing the membership functions that are used.

2

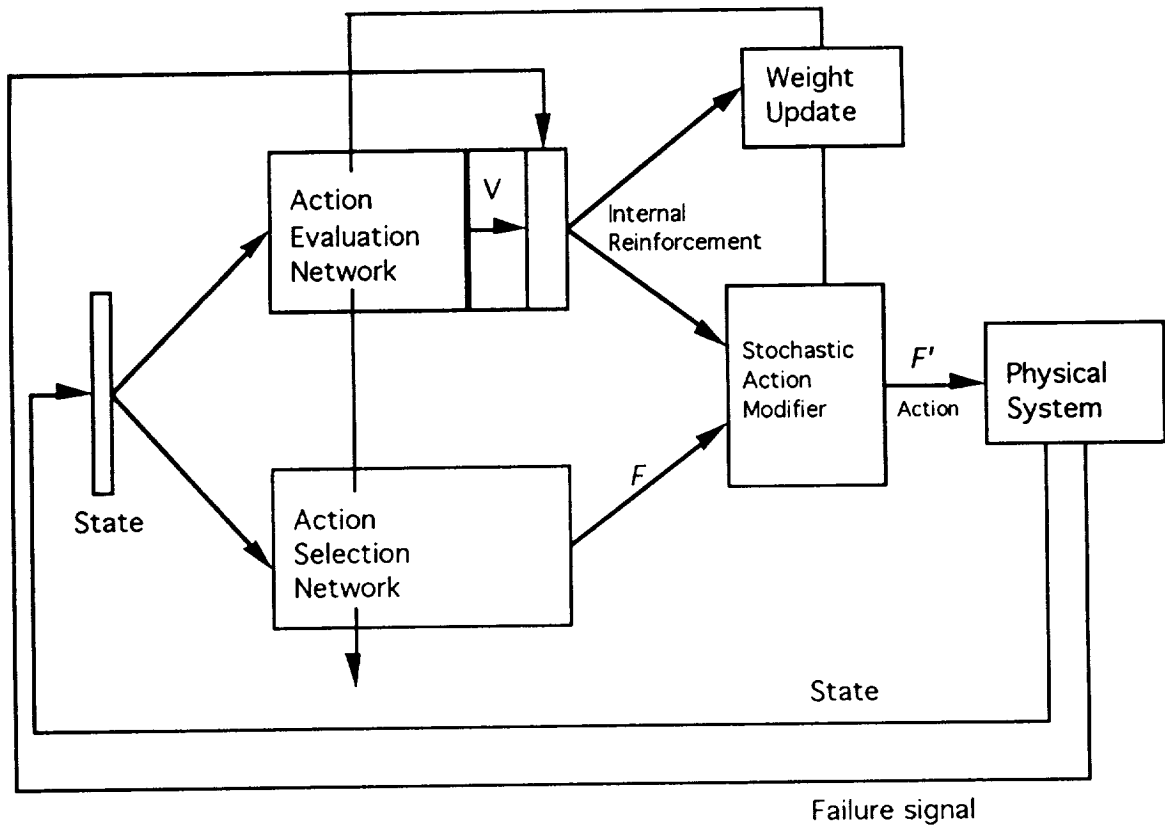


Figure 1: The GARIC Architecture

GARIC is a hybrid architecture for using fuzzy logic control and reinforcement learning. In reinforcement learning, one assumes that there is no supervisor to critically judge the chosen control action at each time step. The learning system is told indirectly about the effect of its chosen control action. GARIC uses reinforcements from the environment to refine its definition of fuzzy labels globally in all the rules and allows any type of differentiable membership function to be used in the construction of a fuzzy logic controller.

The architecture of GARIC is schematically shown in Figure 1. It has three components:

1. The Action Selection Network (ASN) which, given a situation (i.e., a state vector) and by consulting its fuzzy rules, recommends performing a control action F
2. The Action Evaluation Network (AEN) maps a state vector and a failure signal into a scalar score (V) which indicates state goodness. This is also used to produce internal reinforcement.
3. The Stochastic Action Modifier (SAM) uses both F and internal reinforcement to produce an action F' which is applied to the plant.

The ensuing state is fed back into the controller, along with a boolean failure signal. Learning occurs by fine-tuning of the free parameters in the two networks : in the AEN, the weights are adjusted; in the ASN, the parameters describing the fuzzy membership functions change. Further details on GARIC are described in [1].

SPACE SHUTTLE IN-ORBIT OPERATION WITH GARIC

The attitude and translational control are important parts of the Space Shuttle in-orbit operations. The attitude controller performs a variety of tasks including:

- (1) attitude hold or maintaining the desired attitude within a small region of the desired value, typically known as a deadband
- (2) attitude maneuver or going from one attitude to another.

Typical controllers based on the phase plane concept have angle errors and rate errors as input values. The output controller value is a command for generating a correcting torque. For the space shuttle, the rotational corrective torques are generated by thrusters by having compensating thrusters fire along a given axis to nullify the input errors. It uses two types of thrusters (two levels of jet thrusts), known as primary and vernier, and operates with two different sets of deadband values. It can perform rate maneuvers in pulse as well as discrete modes. Typical perturbations acting on the system include gravity gradient, aerodynamic torques, and translational burns.

The translational controller also performs a variety of tasks including the R-bar or V-bar approaches (in which the Orbiter moves along the target's radius vector or velocity vector, respectively, with a sequence of "hops"), station-keeping, and flyaround operations. All testing and training is performed using the Orbital Operation Simulator (OOS) which is a high fidelity shuttle simulator and includes the space shuttle Digital Auto-Pilot (DAP) for attitude operation.

A fuzzy logic controller using 31 rules for each axis (pitch, roll, yaw) have been developed [7], [4]. For each rule, seven labels (Negative-Big, Negative-Medium, Negative-Small, Zero, Positive-Small, Positive-Medium, Positive-Big) are used for angle error and angle error rate, and five labels (NM, NS, ZE, PS, PM) are used for jet firing commands. This controller holds the error between a .5 deadband. If a tighter deadband is required, then the membership functions need to be adjusted manually. However, by using the GARIC architecture, the system learns to automatically adjust its membership functions so that the error remains within the new tighter deadband. In a learning experiment, a failure occurs when the value of a state variable goes beyond the desired deadband. Over a number of trials, and by using the fuzzy reinforcement learning, the GARIC architecture learns to control the error to stay within the new deadband. Similar experiments were also performed for translational control including the R-bar approach, V-bar approach, and fly-around operations.

A set of experiments were performed to tune our fuzzy logic controller to perform a new task of keeping the error within a .4 deadband (i.e., -.4 to +.4) for pitch, roll, and yaw. Less than 10 trials were needed to refine the triangular membership functions as used in our fuzzy rules. Once GARIC has completed its training, we take the refined labels and run the controller again with no on-line learning in order to test its behavior. These experiments showed that GARIC can learn to perform a new task within a limited number of trials in a complex environment such as the simulation of the Space Shuttle in-orbit operations. Further details about these experiments can be found in [4].

Since it is relatively simple to translate a fuzzy rule base into a 5-layer neural network as is used in ASN, then it is expected that GARIC can be applied to other domains where fuzzy logic control has been used. For example, fuzzy logic control based applications in consumer products such as appliances, automobiles, and cameras can use GARIC's method in fine-tuning their performances.

CONCLUSION

GARIC provides a general approach for developing intelligent systems. It starts with the available prior knowledge of the experts in the form of fuzzy rules and refines it using the reinforcements obtained while experimenting with the system. As such, this approach generalizes fuzzy logic control and adds an adaptive behavior to it. In this paper, we briefly discussed an application of GARIC in in-orbit operations of the Space Shuttle. However, a general learning technique as developed in GARIC, may be used in many other domains that fuzzy logic control can be used.

REFERENCES

- [1] Berenji, H.R. and P. Khedkar, Learning and Tuning Fuzzy Logic Controllers Through Reinforcements, vol. 3, no.5, IEEE Transactions on Neural Networks, 1992.
- [2] Berenji, H. R., An Architecture for Designing Fuzzy Controllers using Neural Networks, International Journal of Approximate Reasoning, volume 6, no. 2, pp. 267-292, Feb. 1992.
- [3] Berenji, H. R., On the integration of reinforcement learning and approximate reasoning for control, American Control Conference, page 1900-1904, Brighton, England, 1991.
- [4] Berenji, H.R. and R.N Lea and Y. Jani and A. Malkani and J. Hoblit, A Learning Fuzzy Logic Controller for the Space Shuttle's Orbital Operations, AI Research Branch, FIA-93-30, October 1993.
- [5] Jang, J.S. Self-learning fuzzy controllers based on temporal back propagation, IEEE Transactions on Neural Networks, 3(5), 1992.
- [6] Lea, R. and J. Villarreal and Y. Jani and C. Copeland, Learning Characteristics of a Space-Time Neural Network as a Tether Skiprope Observer, North American Fuzzy Information Processing Society, December 1992", pp. 154-165, Puerto Vallarta, Mexico.
- [7] Lea, R. and J. Hoblit and Y. Jani, Performance Comparison Of A Fuzzy Logic Based Attitude Controller With The Shuttle On-orbit Digital Auto Pilot, North American Fuzzy Information Processing Society, May 1991, Columbia, Missouri.
- [8] Lea, R. and M. Togai and J. Teichrow and Y. Jani, Fuzzy Logic Approach to Combined Translational and Rotational Control of a Spacecraft in Proximity of the Space Station, International Fuzzy Set Association Conference, pp. 23-29, 1989.

A FUZZY CLASSIFIER SYSTEM FOR PROCESS CONTROL

C. L. Karr
U. S. Bureau of Mines
Tuscaloosa Research Center
P.O. Box L, University of Alabama Campus
Tuscaloosa, AL 35486-9777

J. C. Phillips
U. S. Bureau of Mines
Tuscaloosa Research Center
P.O. Box L, University of Alabama Campus
Tuscaloosa, AL 35486-9777

ABSTRACT

A fuzzy classifier system that discovers rules for controlling a mathematical model of a pH titration system has been developed by researchers at the U.S. Bureau of Mines (USBM). Fuzzy classifier systems successfully combine the strengths of learning classifier systems and fuzzy logic controllers. Learning classifier systems resemble familiar production rule-based systems, but they represent their IF-THEN rules by strings of characters rather than in the traditional linguistic terms. Fuzzy logic is a tool that allows for the incorporation of abstract concepts into rule based-systems, thereby allowing the rules to resemble the familiar "rules-of-thumb" commonly used by humans when solving difficult process control and reasoning problems. Like learning classifier systems, fuzzy classifier systems employ a genetic algorithm to explore and sample new rules for manipulating the problem environment. Like fuzzy logic controllers, fuzzy classifier systems encapsulate knowledge in the form of production rules. The results presented in this paper demonstrate the ability of fuzzy classifier systems to generate a fuzzy logic-based process control system.

INTRODUCTION

Researchers at the USBM have developed adaptive process control systems that utilize various tools and techniques from the field of artificial intelligence. The most important artificial intelligence tools used in these controllers have been expert systems, fuzzy logic controllers, and genetic algorithms. The quest for innovative, adaptive, and robust process control systems has progressed to the point that a new control system, called a fuzzy classifier system, has been developed that represents a synergism of several artificial intelligence techniques. The fuzzy classifier system was first proposed by Valenzuela-Rendón [1] to map functions of continuous variables. The current paper describes an extension to the work of Valenzuela-Rendón in that the fuzzy classifier system is used to solve a process control problem. The fuzzy classifier system described represents another step toward truly intelligent computer systems that are capable of manipulating complex problem environments because the fuzzy classifier system, in effect, "learns" to manipulate a pH titration problem environment without prior knowledge of an appropriate set of rules. It is worthwhile to note that this step is based on a large amount of prior USBM research.

Initially, USBM researchers developed expert systems for controlling mineral processing systems [2]. These expert systems utilized traditional production rules of the form IF {condition} THEN {action}, wherein the *conditions* and *actions* were described using conventional set theory. Next, fuzzy logic controllers were developed that replaced the conventional set theory employed by expert systems with fuzzy logic or approximate reasoning [3]. Fuzzy logic allows for the use of membership functions to describe or define abstract terms akin to those used in a human's "rule-of-thumb" approach to decision making [4], and the fuzzy logic controllers that resulted were more efficient than their expert system counterparts. Then, USBM researchers developed an innovative adaptive process control system in which genetic algorithms were used to tune the membership functions associated with fuzzy logic controllers [5-6]. Genetic algorithms are search algorithms based on the

mechanics of natural genetics, and they rapidly locate near-optimum solutions to difficult search problems [7]. The combination of fuzzy logic controllers with genetic algorithms marked a major step in achieving truly adaptive process control. Recently, fuzzy logic controllers have been combined with learning classifier systems to produce fuzzy classifier systems which require minimal information about the physical system being manipulated.

A fuzzy classifier system is a rule-based system that incorporates the "rule-of-thumb" approach used in human decision making with the rule discovery capabilities of learning classifier systems. These innovative systems generate both the rules and membership functions that constitute a fuzzy logic controller. Fuzzy classifier systems consist of three main components: (1) a rule and message system, (2) an apportionment of credit system, and (3) a genetic algorithm. The rule and message system is a mechanism by which the fuzzy classifier system interacts with the problem environment. The fuzzy classifier system receives information concerning the condition of the problem environment and takes an action on the environment based on its rule set. In the apportionment of credit system rules compete for the right to take their action on the problem environment, and are rewarded or punished in accordance with their performance. The genetic algorithm is used as a rule discovery system in which new rules are generated and inserted into the rule store of the fuzzy classifier system. The rewards and penalties accrued via the apportionment of credit system drive the genetic algorithm's search.

In this paper, the fuzzy classifier system developed at the USBM is applied to a specific problem environment: a pH titration system. The fuzzy classifier system developed for a computer simulation of the pH system is quite effective in achieving the process control objective. Its performance is compared to that of a fuzzy logic controller that has been shown to manipulate the pH environment in an effective manner. The performance of the fuzzy classifier system demonstrates the potential of this approach to adaptive process control. Its effectiveness in the highly nonlinear problem of pH control points to far-reaching implications for application in diverse industrial fields such as mineral processing, chemical engineering, and solid waste disposal.

The fuzzy classifier system that results from the synergism of artificial intelligence techniques is important to a number of industries for two reasons: (1) it is simple and (2) it is versatile. First, the use of fuzzy logic greatly simplifies the task of developing a rule-based controller, and rule-based controllers have been successfully implemented in numerous problem domains. Second, the approach to process control is versatile because it relies on genetic algorithms which have been used to efficiently solve a wide spectrum of search problems. Furthermore, the basic design of the control systems that result allow for the effective manipulation of complex problem environments despite the fact that the control systems have not been provided with rules for manipulating the problem environment. This trait is a virtual necessity in numerous industrial settings such as the minerals industry in which the mechanics of the processes in the plants are quite complicated and not understood well enough to write complete rule sets.

THE PHYSICAL SYSTEM

A simple laboratory pH system is considered to present the details of a fuzzy classifier system. A schematic of this pH system is shown in Figure 1. The system consists of a tank initially containing a given volume of a solution having a known pH. There are two valved input streams into the beaker. The valves on these two *control input streams*, one a strong acid (0.1 M HCl) and one a strong base (0.1 M NaOH), can be adjusted to cause a change in the pH of the solution in the tank. The objective of the control problem is to neutralize the solution — drive the pH to 7 — in the shortest time possible by adjusting the valves on the control input streams. Additionally, the valves on the control input streams are to be fully closed after the solution is neutralized. As a constraint on the control problem, the valves can only be adjusted a limited amount (0.5 mL/s, which is 20 pct of the maximum flow rate of 2.5 mL/s), to restrict pressure transients in the associated pumping systems.

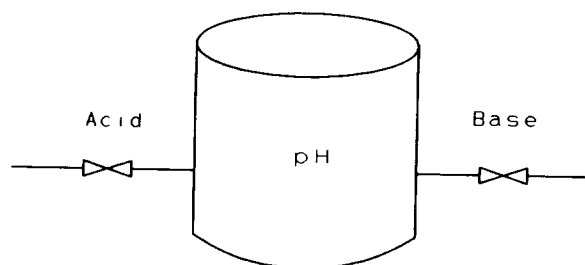


Figure 1. Schematic of the pH titration system

The development of a fuzzy classifier system requires a computer model of the problem environment, in this case the pH titration system. A model of the pH system is required due to the way in which a learning portion of the fuzzy classifier system operates. The fuzzy classifier system employs a genetic algorithm which evaluates a number of possible solutions to the problem of locating rules that are appropriate for completing the titration. Some of the possible rule sets the genetic algorithm investigates are totally unacceptable: they represent preposterous control strategies. Therefore, the potential solutions are investigated on a computer simulation of the pH system. Fortunately, the dynamics of the pH system are well understood, and can be modeled for a variety of reactions using conventional techniques [8]. In the pH system considered here, the development of a model of the physical system does not present an insurmountable obstacle. However, it should be realized that for many complex industrial systems, the development of an accurate computer model is an imposing task. In these cases some form of empirical model, such as a statistical or neural network model, offers a suitable alternative to a first principle model. It should also be noted that in situations such as the pH system in which a fundamental model can be produced, conventional control strategies can be quite effective. However, some systems are modeled using statistical approaches (even neural networks and fuzzy rule bases), and it is in these situations that the fuzzy classifier system has the greatest potential.

LEARNING CLASSIFIER SYSTEMS

A classifier system is a genetics based machine learning system that learns rules, called classifiers, to govern its performance in a given environment e.g., the pH titration system. The systems are based on a model of a service economy in which money is exchanged for services, and include a mechanism for discovering new rules. Unlike in traditional expert systems, a rule's relative value is learned, not fixed by the programmer. In classifier systems the rules are forced to coexist in a service economy where a competition is held to decide which rule will be put into effect under a specified set of conditions in the environment. The competitive nature of the economy ensures good rules survive and bad rules die off, while the exploratory portion of the learning classifier system creates new rules. These systems operate incrementally, testing new rules while steadily improving performance in an environment.

In general, classifier systems are composed of three subsystems:

- 1) Rule and message system;
- 2) Credit assignment system;
- 3) Rule discovery system.

These three subsystems, when combined with a mathematical model of the environment, form a computationally complete system capable of learning effective rules for interacting in an environment and controlling processes. The mathematical model offers an arena in which the fuzzy classifier system can investigate new and improved control strategies. Furthermore, this model of the problem environment can be used to compute changes in the problem environment that can not be measured directly [6].

The rule and message system is similar to a production system. Production systems are schemes that use rules as their only means of operation. The rules are generally of the form:

IF {condition} THEN {action}.

When the *condition* exists in the environment, then the *action* is to be taken. In learning classifier systems the rule and message system involves the completion of some basic tasks. The learning classifier system evaluates the existing state of the environment; it determines what conditions exist. These conditions are compared to the current rule set to determine which rules are eligible to be put into effect. A competition is held among the eligible rules and a set of winning rules is selected. These rules take their associated action causing a change in the environment. It should be noted that the first competition is strictly a random decision since all rules initially have the same relative strength. However, as will be seen shortly, all future competitions are decided based upon the previous successes or failures of individual rules. At this point the rule and message system becomes temporarily inactive and the credit assignment system takes over. Before the credit assignment system is discussed, note that the rule and message system is designed to activate several classifiers in parallel which actually characterizes learning classifier systems as parallel production systems even though they are readily implemented on sequential machines.

The purpose of the credit assignment system [9] is to evaluate how useful each classifier has been in producing desirable effects on the environment (how useful it has been in solving the problem at hand). The method used most often is the bucket brigade algorithm of Holland [10]. In the bucket brigade algorithm every classifier is assigned a strength which is a value representing that classifier's usefulness in solving the problem. The classifiers that are eligible to take an action at a given time step bid a portion of their associated strength for the right to take their action. Once all the eligible rules have made their bids, winners are selected through probabilistic means based on the size of the bids; those who bid highest have the highest chance of being selected to take their action. If the environment is in a more desirable state than at the previous time step, the auction winners pay their bids to the classifiers that took action at the previous time step. Thus, the rules that caused desirable changes in the environment are rewarded with payoff. A mandatory payment is also received from every classifier at each time step. This payment, or tax, helps weed out poor rules because in the rule discovery system a classifier's existence depends on its associated strength. If a classifier fails to win the auction and cause a positive change on the environment, its chance of survival is reduced. The bucket brigade algorithm's redistribution of classifier strengths leads to bids that are representative of classifiers' potential for achieving the goal; effective rules bid proportionately higher.

The purpose of the rule discovery system is to generate new rules with potential for more efficiently reaching the goal assigned to the learning classifier system. The genetic algorithm is the technique generally used in rule discovery systems. The genetic algorithm requires the elements of the search space (in this case the rules) to be coded as finite length strings. Effective classifiers, those with high associated strengths, are selected for combination with other highly fit classifiers. Portions of the classifiers are selected at random and combined with portions of other classifiers through the standard genetic algorithm operators of reproduction, crossover, and mutation to produce new rules. The intent is to combine advantageous qualities from two separate classifiers to form two new, more highly fit classifiers.

The three subsystems discussed in this section form a system capable of learning and evaluating new rules for interacting in an environment. A more detailed description of learning classifier systems can be found in Goldberg [7].

FUZZY LOGIC CONTROLLERS

The popularity of fuzzy logic controllers has increased dramatically in the last decade. This increase in popularity has also provided an increase in the number of approaches to implementing fuzzy logic in a process control system [11]. Perhaps the most efficient way to introduce fuzzy logic controllers is to provide

a step-by-step procedure that can be used. In this section, such a procedure is presented and applied to the pH titration system. The fuzzy logic controller that results is later used to evaluate the performance of a fuzzy classifier system.

The first step in developing a pH fuzzy logic controller is to decide on the *condition variables* (these variables appear on the left side of the fuzzy logic controller rules which are of the form: **IF {condition} THEN {action}**). Certainly there are numerous condition variables that could be considered in the pH system (pH of solution in the tank, flow rates of the input streams, concentrations of input solutions, volume in the tank, and many others). However, it is important to limit the number of condition variables used to a small fundamental set because the size of the rule set increases multiplicatively with the number of condition variables. After a period of experimentation (an inevitable requirement for the development of a quality fuzzy logic controller), two condition variables were selected: the current value of pH (pH) in the beaker and the current time rate of change of the pH in the tank (ΔpH).

The second step is to determine the specific actions that can be taken on the system, i.e., the *action variables* must be determined. In the pH system, the determination of the action variables is relatively straightforward. There are basically only two action variables that can be altered by the controller: the valve settings (and thus the flow rates) associated with the control input streams. Therefore, the two action variables are the flow rates for the strong acid (Q_{ACID}) and the strong base (Q_{BASE}), respectively, of the input streams. The selection of the action variables differs from the selection of the condition variables in that the number of action variables has no effect on the number of rules required.

The third step is to choose linguistic terms that represent each of the condition and action variables. Eight terms were used to describe pH, four terms were used to describe ΔpH , and four terms were used to describe both Q_{ACID} and Q_{BASE} . The specific linguistic terms used to describe the pertinent variables in the pH system follow:

pH	Very Acidic (VA), Acidic (A), Mildly Acidic (MA), Neutral Acidic (NA), Neutral Basic (NB), Mildly Basic (MB), Basic (B), and Very Basic (VB);
ΔpH	Negative Large (NL), Negative Small (NS), Positive Small (PS) and Positive Large (PL);
Q_{ACID}	Zero (Z), Small (S), Medium (M), and Large (L);
Q_{BASE}	Zero (Z), Small (S), Medium (M), and Large (L).

All of these linguistic terms are subjective, i.e., the terms can mean different things to different people, but the developers (the authors) of the pH fuzzy logic controller have some meaning they associate with each of the terms.

The fourth step is to provide the selected linguistic terms with some concrete, or crisp meaning. The linguistic terms are "defined" by membership functions. As with the requirement for selecting the necessary linguistic terms, there are no definite guidelines for constructing the membership functions; the terms are defined to represent the designers' general understanding of what the terms mean.

The fifth step in the design of a fuzzy logic controller is the development of a rule set. The rule set in a fuzzy logic controller must include a rule for every possible combination of the controlled variables as they are described by the chosen linguistic terms. Thus, the pH fuzzy logic controller, as described to this point, will contain $8 * 4 = 32$ rules to describe all of the possible conditions that could exist in the pH system as described by the linguistic terms represented by the membership functions. For any combination of the condition variables, an appropriate choice of the action variables is prescribed. Due to the nature of the linguistic terms, most of the actions needed for the 32 possible condition combinations are readily apparent. For instance, when pH is VA and ΔpH is NS, then Q_{ACID} should be Z and Q_{BASE} should be L. However, there are some conditions for which the appropriate action is not readily apparent. In these instances, some experimentation is often needed.

Now that both the condition and action variables have been chosen and described with linguistic terms, and a rule set has been written that prescribes an appropriate action for every possible set of conditions, it is left to determine a single crisp value of the acid and base valve settings. This may or may not be viewed as a step in the fuzzy logic controller development. Certainly, there are numerous approaches to the fuzzy computations performed by the fuzzy logic controller. The procedure for determining a single crisp value of the valve settings for the acid and base input streams is a concern because, unlike in traditional expert systems, more than one of the fuzzy logic controller's 32 rules can be applicable for a given state of the pH system. A common technique for accomplishing this task is the center of area method (sometimes called the centroid method). In the center of area method, the action prescribed by each rule plays a part in the final crisp value of the valve settings. The contribution of each rule to the final value of Q_{ACID} and Q_{BASE} is proportional to the minimum confidence (the minimum value of the membership function values on the left side of the rule) one has in that rule for the specific state of the physical system at the particular time. This is equivalent to taking a weighted average of the prescribed actions. Further explanation of the center of area method is provided in Sugeno [11].

One detail specific to the pH system should be considered here. There is a limit on the allowable change in the flow rates of the input streams, i.e., the flow rates cannot change by more than 0.5 mL/s. However, the membership functions used in the center of area method (shown in Figure 2) allow for values of Q_{ACID} and Q_{BASE} to range between 0.0 mL/s and 2.5 mL/s, irrespective of their current values. The constraint is imposed by computing the value of the flow rates using the center of area method. If this value exceeds the constrained flow rate (for a time period between steps of 1 s), the flow rate is changed by the maximum allowable value of 0.5 mL/s (for either increases or decreases in flow rate). With the determination of a strategy for resolving "conflicts" in the actions prescribed by the individual rules, the fuzzy logic controller is complete.

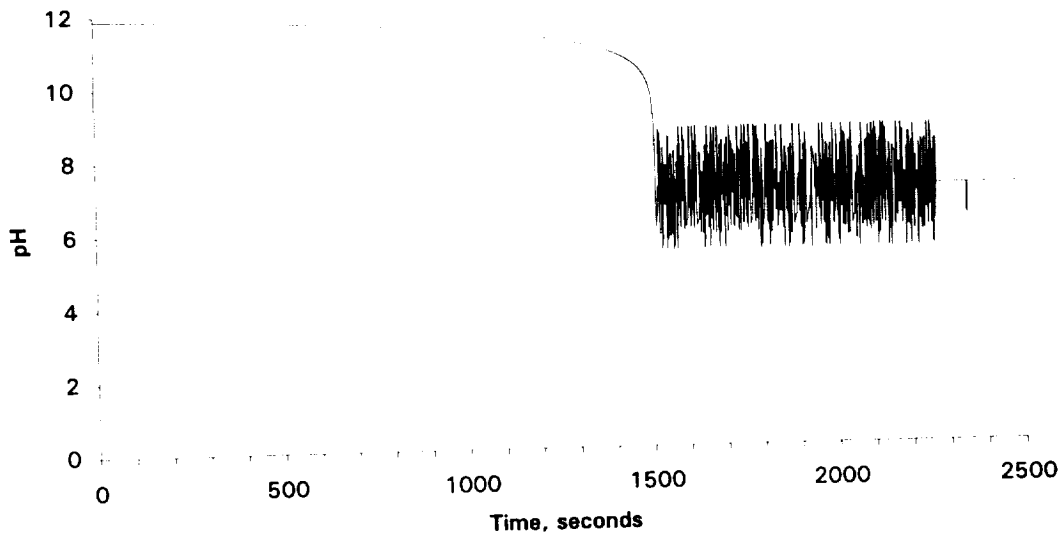


Figure 2. The fuzzy classifier system neutralizes a basic solution

APPLICATION OF A FUZZY CLASSIFIER SYSTEM

The first step in producing a fuzzy classifier system for the pH titration system is to develop a means of representing linguistic rules and fuzzy membership functions as strings of characters. This step is required so that a genetic algorithm may be employed in the rule discovery system. An approach to mapping fuzzy membership functions has been well documented [5] and is used here to form a portion of the character strings. The remainder of the strings represent the rule set used in the controller. For this study a direct mapping was used to represent the rule sets. The condition portion of each classifier is a function of the current pH and the current

time rate of change of pH. The action portion of each classifier involves adjusting the net inflows (Q_{ACID} and Q_{BASE}) at the following time step. The coding used for the rule set is summarized in Table 1. It is important to note that Table 1 includes only the coding used for the rules. The character strings used contained information concerning both the rules and the membership functions and were 69 bits long.

Table 1. - The coding used to represent the rule sets.

Condition				Action			
pH		ΔpH		Q_{ACID}		Q_{BASE}	
binary	set	binary	set	binary	set	binary	set
000	VA	00	NL	00	Z	00	Z
001	A	01	NS	01	S	01	S
010	MA	10	PS	10	M	10	M
011	NA	11	PL	11	L	11	L
100	NB						
101	MB						
110	B						
111	VB						

The basic binary coding in which each bit position is characterized by either a 1 or a 0 would be sufficient for the learning classifier system. However, to add flexibility and allow for more general rule representation, a third character is introduced, namely the # or "don't care" symbol. A bit position filled by a # is matched by either a 1 or a 0. An example rule follows:

condition → action
 00#00 → 0011.

In this rule, the value of pH at the current time is defined by the bits 00#. This condition is satisfied by either 000 or 001. Thus, this rule says that if the pH is currently very acidic or acidic with a negative large time rate of change of pH, then minimize the flow of acid into the tank and maximize the flow of base into the tank at the next time.

Now that a means for representing the rules and fuzzy membership functions has been established, the implementation of a fuzzy classifier system is discussed. A rule set of 250 rules was used to control a mathematical model of the pH titration system. These rules were used to manipulate the computer model of the pH titration system until 25,000 rules had been enacted. At the end of this period, 120 new rules were produced using a genetic algorithm. These 120 rules replaced 120 of the 200 poorest rules in the previous rule set. By crowding out the poor rules and relying on the genetic algorithm's survival-of-the-fittest approach, the rule sets achieved higher levels of performance. A learning cycle consisted of the 25,000 actions or one generation of new rules using the genetic algorithm. However, learning actually took place after each individual rule activation at which time the bucket brigade algorithm redistributed payment to the rules. The steps composing these learning-cycles are relatively straightforward and consist of the following:

- 1) Evaluate the condition or state of the physical system at the current time step as predicted by the model.
- 2) Compare the state of the physical system to the condition portion of each of the 250 classifiers to determine which rules are eligible to take their action.

- 3) Hold an auction between the eligible rules to determine which rules get to take their actions. The bid placed by a given rule is defined by

$$BID = S * C_{bid} \quad (1)$$

where BID is the rule's bid, S is the rule's strength, and C bid is a constant. In this work, C is set to a value of 0.1 which places an appropriate emphasis on the bid. Within the model of a service economy, strength becomes a measure of how effective a rule is at driving the system to its setpoint.

- 4) Evaluate the effectiveness of the action. If the action, which is a function of both the rules and the fuzzy membership functions, drives the system closer to its setpoint, then the rules that won the auction at the previous time step are rewarded with the current winners' bids and a tax paid by each classifier as described below. Otherwise no rule receives a reward and the tax is accumulated in a fund and carried over to the next time step.

- 5) Tax each of the rules in the current rule set. Every rule is taxed according to the following:

$$TAX = (C_{bidtax} - C_{lifetax} S) \quad (2)$$

where TAX is the tax paid by a classifier, C_{bidtax} is a constant (equal to $5 \cdot 10^{-5}$ in this study if the rule's condition is met and thus it placed a bid; it is equal to 0.0 otherwise), and $C_{lifetax}$ is a constant (equal to $2 \cdot 10^{-5}$ in this study under every circumstance).

- 6) After 25,000 actions are taken, apply the genetic algorithm to generate a new rule set which includes 120 new rules. The genetic algorithm combines portions of the most effective rules as determined by their strength, which rises if a rule is effective and falls if a rule is non-effective due to the nature of the apportionment of credit algorithm.

Although the mechanics are relatively simple, the fuzzy classifier system is able to employ the exploratory power of the genetic algorithm and the structured credit assignment of the bucket brigade to learn new and improved rules for controlling pH titration system.

RESULTS

Evaluating the effectiveness of a process control system is not a trivial endeavor; there are various criteria for efficient control that can be established, and there are numerous conditions over which the controller must be able to perform. In evaluating the performance of the fuzzy classifier system it is important to keep in mind the control objective which is to neutralize the solution as fast as possible while not violating the constraints placed on the flow rates. Additionally, it is important to realize that if the controller can accomplish the control goal from extreme portions of the control space (when the solution is initially extremely acidic or extremely basic), it should perform well in the more moderate portions of the control space. Thus, the performance of the fuzzy classifier system is compared to the performance of a fuzzy logic controller that was developed by the authors for the pH titration system.

Figure 2 summarizes the performance of the fuzzy classifier system. In the particular case depicted, the fuzzy classifier system was posed with the problem of neutralizing a basic solution. As can be seen in the figure, the fuzzy classifier system rapidly locates rules for driving the pH to values that range roughly between 5.2 and 8.5. Then, after a brief period of exploration, the controller is able to evolve into a form that forces the pH to the desired value of 7. During the period of exploration, the fuzzy classifier is considering rules selected by a genetic algorithm; some are good and some are bad. The result appears to be a random walk through the search space. However, the genetic algorithm is actually updating its selections based on the performance of the rules. Figure 3 shows the performance of a fuzzy logic controller that was designed by the authors to solve the pH problem. Note that this controller drives the pH to a value that is between 6.5 and 7.5. It is important to note that the

performance of the author-developed fuzzy logic controller can be improved by altering the membership functions [6], but this is a time-consuming task to accomplish without the assistance of some computational algorithm such as a genetic algorithm. Nonetheless, the two figures demonstrate the fact that the fuzzy classifier system contains a mechanism for improving its performance through the discovery aspects of a genetic algorithm.

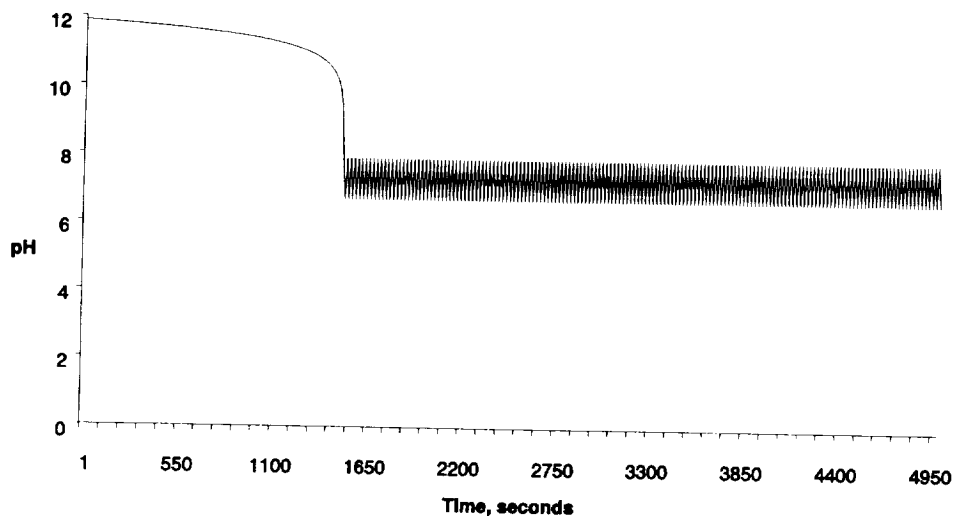


Figure 3. A fuzzy logic controller neutralizes a basic solution

Figures 2 and 3 demonstrate the effectiveness of the fuzzy classifier system in the pH problem environment. Although the results presented do not alleviate concerns as to the ability of the fuzzy classifier system to perform effectively in process control problems, they do point to the potential of these fuzzy rule discovery systems. These results demonstrate the ability of the system to discover both rules and fuzzy membership functions for effectively manipulating a highly nonlinear physical environment.

FUTURE WORK

The previous section provided results that indicate a fuzzy classifier system is capable of efficiently controlling a pH titration system. However, there are still a number of research issues left open. The following amplifications are currently under investigation:

- The pH titration system is being extended to include external perturbations such as additions of a buffering solution. This extension will make the pH titration system more like chemical systems currently used in industry, and will force the fuzzy classifier system to discover rules in real time more often than it has to with the current physical system.
- A general purpose credit assignment system is being developed. Currently, the credit assignment system is specialized for the pH titration system. For the fuzzy classifier system to be easily applied to alternative process control problems, a general purpose credit assignment algorithm must be developed.
- The fuzzy logic control aspect of the fuzzy classifier system is being improved. The center of area algorithm for selecting a single crisp action is but one of a number of potential solution algorithms that have been proven effective in fuzzy logic controllers. Current efforts are centered on including fuzzy singletons [11] into the fuzzy classifier system. Successful implementation of this algorithm should improve the effectiveness of the fuzzy actions prescribed by the rule sets.

- Alternative schemes for encoding the classifiers are being considered. The effectiveness of the genetic algorithm's search is highly dependant on the coding scheme used. Although the current coding seems to be effective, recent research efforts [5] indicate that there may be more efficient ways to encode the information concerning the rules and the fuzzy membership functions.
- A micro genetic algorithm is being incorporated into the fuzzy classifier system. The micro genetic algorithm is a small population genetic algorithm that has been reported to perform well across a spectrum of search problems, and in the problem of locating fuzzy membership functions in particular [6]. If this effort is successful, the fuzzy classifier system should be able to locate effective rules in less time than it currently takes.

REFERENCES

1. Valenzuela-Rendón, M. (1991). "The Fuzzy Classifier System: A Classifier System for Continuously Varying Variables." In K. Belew and L. Booker (Eds.), *Proceedings of the Fourth International Conference on Genetic Algorithms*, 346-353.
2. Davis, B. E., Jordan, C. E., and Stanley, D. A. (1990). "Expert Advisor for Phosphate Flotation." *Control '90 - Society for Mining, Metallurgy, and Exploration*, 77-85.
3. Karr, C. L., Freeman, L. M., & Meredith, D. L. (1990). "Genetic Algorithm Based Fuzzy Control of Spacecraft Autonomous Rendezvous." *Proceedings of Fifth Annual Conference on Artificial Intelligence for Space Applications*, NASA CP-3073, 43-51.
4. Zadeh, L. A. (1973). "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes." *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, 28-44.
5. Karr, C. L. (1991). "Genetic Algorithms for Fuzzy Controllers." *AI Expert*, 6(2), 26-33.
6. Karr, C. L., and Gentry, E. J. (1993). "Fuzzy Control of pH Using Genetic Algorithms." *IEEE Proceedings on Fuzzy Systems*, 1(1), 46-53.
7. Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.
8. Hand, C. W., and Blewitt, G. L. (1986). *Acid-Base Chemistry*. Macmillan Publishing Company, New York, NY.
9. Minsky, M. L. (1967). *Computation: Finite and Infinite Machines*, Prentice Hall, Englewood Cliffs.
10. Holland, J. H. (1962). "Outline for a Logical Theory of Adaptive Systems." *Journal of the Association of Computing Machinery*, 3, 297-314.
11. Sugeno, M. (Ed.). (1985). *Industrial Applications of Fuzzy Control*. Elsevier Science Publishers, Amsterdam.

**FUZZY-NEURAL CONTROL OF AN
AIRCRAFT TRACKING CAMERA PLATFORM**

Dennis McGrath
Naval Air Warfare Center - Aircraft Division
Lakehurst, New Jersey

ABSTRACT

A fuzzy-neural control system simulation was developed for the control of a camera platform used to observe aircraft on final approach to an aircraft carrier. The fuzzy-neural approach to control combines the structure of a fuzzy knowledge base with a supervised neural network's ability to adapt and improve. The performance characteristics of this hybrid system were compared to those of a fuzzy system and a neural network system developed independently to determine if the fusion of these two technologies offers any advantage over the use of one or the other. The results of this study indicate that the fuzzy-neural approach to control offers some advantages over either fuzzy or neural control alone.

INTRODUCTION

Intelligent control has been dramatically affected in recent years by both fuzzy logic systems and artificial neural networks. Fuzzy logic controllers (FLC) provide a method for encoding knowledge of a system in non-exact (fuzzy) terms. The performance of fuzzy control systems with several simple rules has been successfully applied to a variety of non-linear control tasks. Unfortunately, the tuning required for conventional fuzzy systems to achieve maximum performance is often a time consuming trial-and-error process. Alternatively, neural networks (NN) provide a less structured, self-adaptive method for control where knowledge is acquired from experience. The effectiveness of these networks, however, is dependent on the quality and quantity of the training data. Many researchers have recently proposed methods for combining fuzzy logic with neural networks. The fusion of these two technologies has yielded a new type of intelligent control which exploits the advantages of both technologies, allowing the control engineer to combine knowledge *a priori* with knowledge *a posteriori*.

A camera mounted on a positioning platform can be used in aircraft guidance and recognition during final approach to an aircraft carrier. The movement of such a platform is dependent on relative position error and range of the target, as shown in Figure 1. The problem of controlling this platform is well-suited for fuzzy control, since a few simple rules can describe the desired characteristics of the system. Previous studies, which have compared fuzzy tracking to traditional target tracking methods such as the Kalman Filter method, have shown that FLC trackers can match or surpass the performance of conventional tracking systems[2].

The success of previous research in this area inspired the idea that fuzzy target tracking could be improved through some adaptive learning method. In this project, an aircraft tracking camera platform simulation was developed using a FLC tracking system. For the sake of comparison, a simple backpropagation network was implemented for the same task. A fuzzy neural network (FNN) was then

developed using the fuzzy-neural cooperative method proposed by Kawamura et. al.[1]. This method was explored as a means of achieving adaptive fuzzy control by combining the structure of a fuzzy logic control system with the adaptive learning capabilities offered by neural networks.

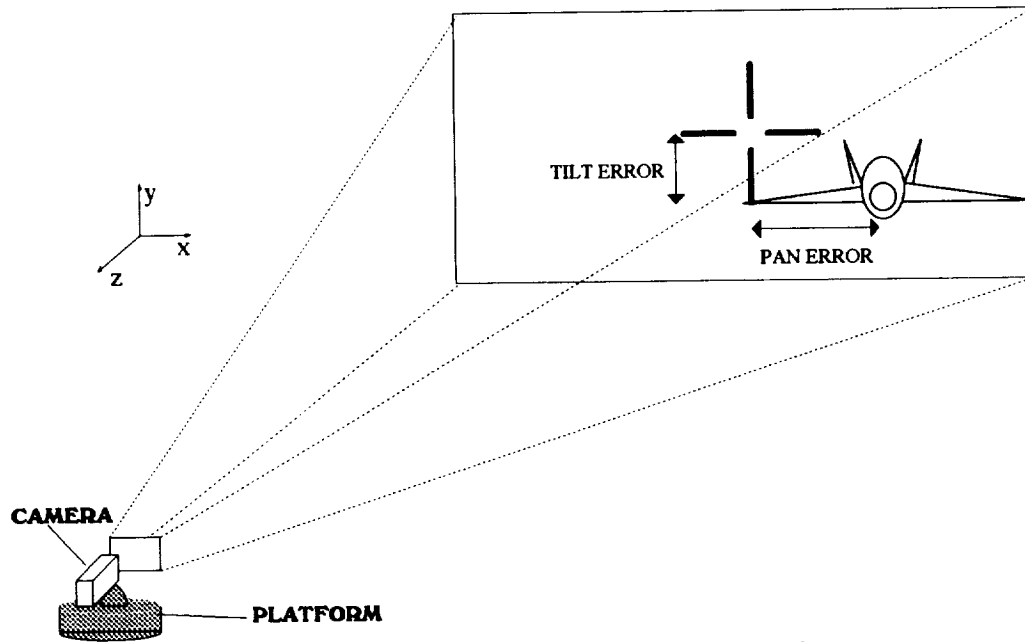


Figure 1 - Aircraft Tracking Using a Positioning Platform

FLC and NN TRACKING SYSTEMS

In camera tracking, the objective is to maintain a target image in the center of the camera's field-of-view (FOV). Typically, the exact x, y, and z location of the target is not available to the system[3]. The position error must be measured in pixels as the difference between the center of the camera's FOV and the centroid of the target image. Likewise, range can only be determined by the relative size (number of pixels) of the target image. For this simulation, however, explicit position information was available from radar observations of several approaches. The data was sampled at 20 Hz, and the approaches typically lasted 45 seconds giving approximately 900 data points per approach.

Pan and tilt are independent functions controlled by separate servos. The rates of pan and tilt are dependent on azimuth and elevation error, respectively, as well as range. In this simulation, a maximum pan rate of $\pm 15^\circ/\text{s}$ and a maximum tilt rate of $\pm 5^\circ/\text{s}$ were used. In the case of carrier approaches, panning is a much more difficult tracking task than tilting, since more abrupt aircraft movements take place in the x domain than in the y domain. Therefore, only the pan functions will be described for the remainder of this paper, with the understanding that the tilt control system is functionally identical to the pan system.

Fuzzy rules for camera tracking are easily derived by verbalizing the method that a person might use to track a moving target in photography or skeet shooting. These rules take the form of "if the target is to the left and it is close, then pan quickly to the left...", and so on. The FLC tracking system developed for this study is shown in Figure 2. It employs three membership functions for ERROR and two for RANGE. Six rules correspond to five output membership functions. Centroid defuzzification is used to determine the value of PANRATE.

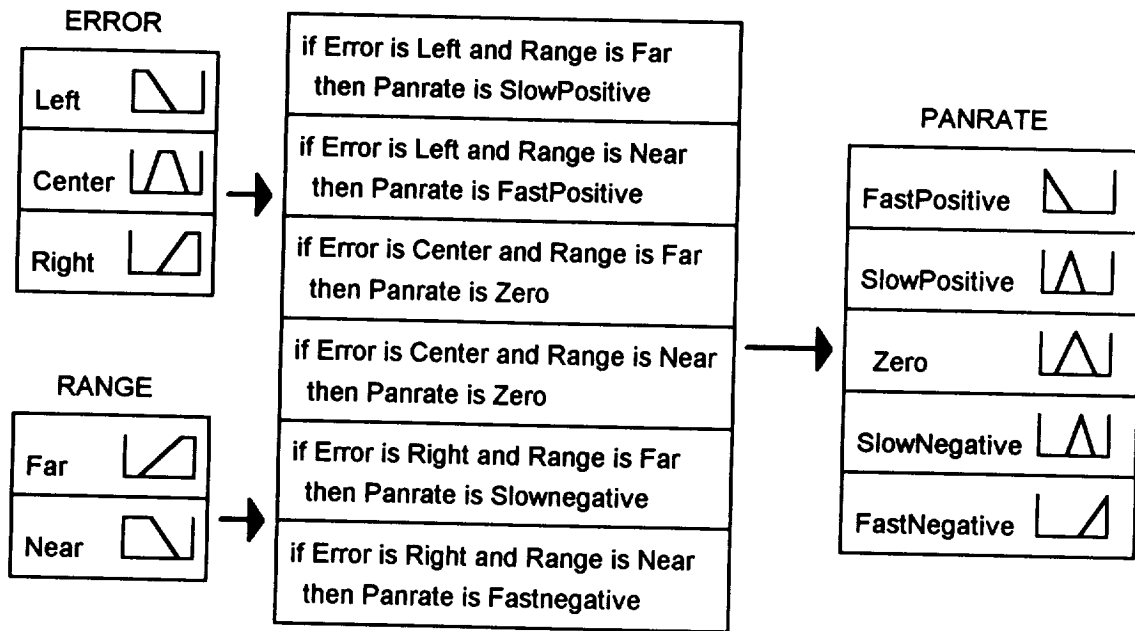


Figure 2 - A fuzzy logic controller for aircraft tracking

The same functions can be performed by a backpropagation network consisting of two inputs, one output, and one hidden layer as shown in Figure 3.

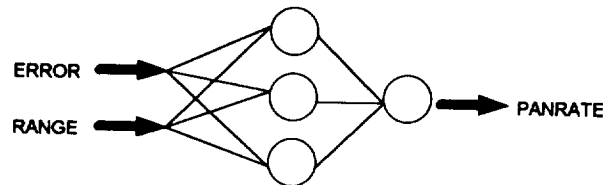


Figure 3 - A simple backpropagation network for aircraft tracking

3. FNN TRACKING SYSTEM

The development of a FNN tracking system required that a network be created which modeled the behavior of the fuzzy tracking system. In this way, the essential operations of fuzzification, inferencing, and defuzzification are accomplished through a series of functions represented by selectively connected nodes. The network architecture and initial system parameters must be empirically developed so that the FNN matches the functions of the existing FLC.

Input membership functions are easily realized through the use of the sigmoidal functions associated with neurons:

$$S(\sum w_{ij}o_i) = \frac{1}{1 + e^{-\sum w_{ij}o_i + \theta}} \quad (1)$$

where $\sum w_{ij}o_i$ is the sum of the input-weight products and θ is the bias or threshold. As Figure 4 shows, the input functions for ERROR and RANGE are represented as logistic sigmoids for Left, Right, Near, and Far and as the difference between two sigmoids in the case of the bell-shaped Center function. The appropriate weight and threshold values were determined, and the antecedent membership functions take the following form:

$$LEFT = \frac{1}{1 + e^{13.3ERROR - 3.3}} \quad (2a)$$

$$CENTER = \frac{1}{1 + e^{-13.3ERROR + 3.3}} - \frac{1}{1 + e^{-13.3ERROR + 9.9}} \quad (2b)$$

$$RIGHT = \frac{1}{1 + e^{-13.3ERROR - 9.9}} \quad (2c)$$

$$FAR = \frac{1}{1 + e^{-13.3RANGE - 6.6}} \quad (2d)$$

$$NEAR = \frac{1}{1 + e^{13.3RANGE - 6.6}} \quad (2e)$$

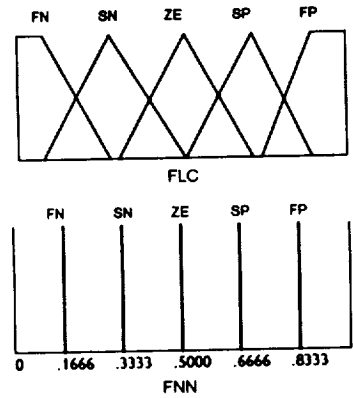
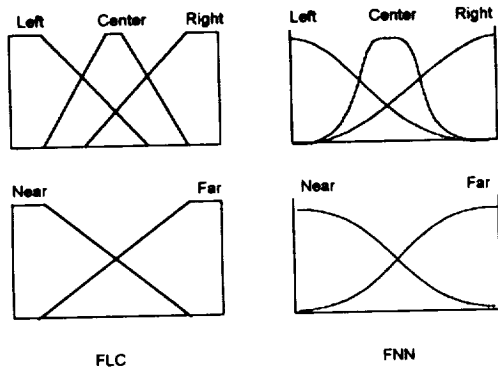


Figure 4 - Antecedent membership functions as sigmoids **Figure 5 - Consequent membership functions as singletons**

The inference process correlates input membership values to output membership values. All rules in this system are conjunctive, meaning that they feature the AND operation. Using max-product inferencing, the result of the AND operation is simply the product of the inputs. Alternatively, max-min inferencing could be used, where the conjunctive inference nodes would use the $\min(A,B)$ operator.

Output (consequent) membership functions are represented as singleton functions for the purpose of simplifying the defuzzification process. Singleton functions have only one value. In this case they represent the approximate centroid of the original output membership functions, as shown in Figure 5. By assigning the singleton values to weights at the output layer, defuzzified values can be computed as a weighted average of these singletons based on the truth value of the consequent membership :

$$OUTPUT = \frac{\sum O_i w_{ij}}{\sum O_i} \quad (3)$$

The resulting network takes the form shown in Figure 6. Hidden layers represent input, rules, and output functions. When ERROR and RANGE values, normalized to [0,1], are fed forward into the network, they are converted to membership values. These membership values are then processed by rule nodes, yielding output (consequent) membership values. These membership values are then defuzzified as explained previously. To this point, the behavior of the FNN is identical to that of the FLC. Parametric learning is then accomplished via the generalized delta rule, as in many other backpropagation

networks.[4] The system error is simply the difference between the new center of the camera's field-of-view and the actual aircraft position. Adjustments are made to weights at the input and output layers based on the calculated deltas derived from the system error, according to the relation:

$$\Delta w_{jk} = \eta \delta_k o_j \quad (4)$$

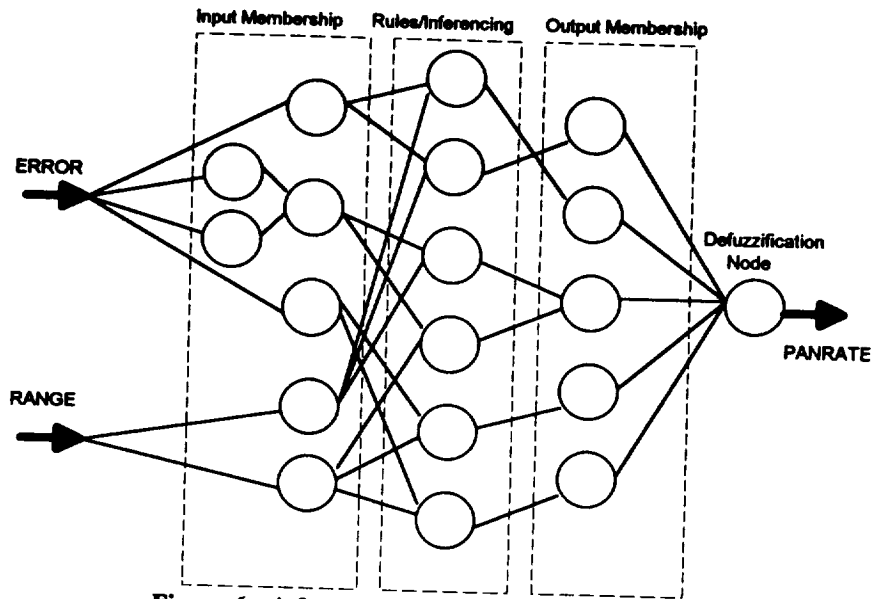


Figure 6 - A fuzzy neural network for aircraft tracking

SIMULATION RESULTS

Using data from several different carrier approaches, the FLC, NN, and FNN tracking methods were implemented in simulation. Figure 7 is a comparison of the performance of each system on a section of the test approach after one, fifteen, and fifty approaches. The dotted lines represent the aircraft movement about the centerline of the approach path, and the solid lines represent the line of sight of the camera.

The performance of the FLC was good, but the camera's line-of-sight lagged slightly behind the aircraft's position. Since it has no parametric adjustment capability, its performance obviously did not improve with time. As was expected, the NN tracker began with random behavior on the first approach. By the end of the first approach, the network began to learn the basics of panning and tilting, but left much room for improvement. The performance had improved substantially by the fifteenth approach, where the line-of-sight lagged slightly and tended to overshoot the aircraft during motion reversals. By the fiftieth approach, the lag had diminished to near-zero, but some overshoot was still evident.

The FNN, even on the first approach showed improvement over the FLC, demonstrating much less lag and a slight overshoot. By the fifteenth approach, the lag was nearly eliminated and the overshoot was significantly reduced. After the fiftieth approach, the line-of-sight of the FNN tracker was virtually indistinguishable from the path of the aircraft. Since it began with the intrinsic knowledge of the FLC, its learning outpaced that of the NN tracker. Under all conditions, the FNN outperformed the other two methods.

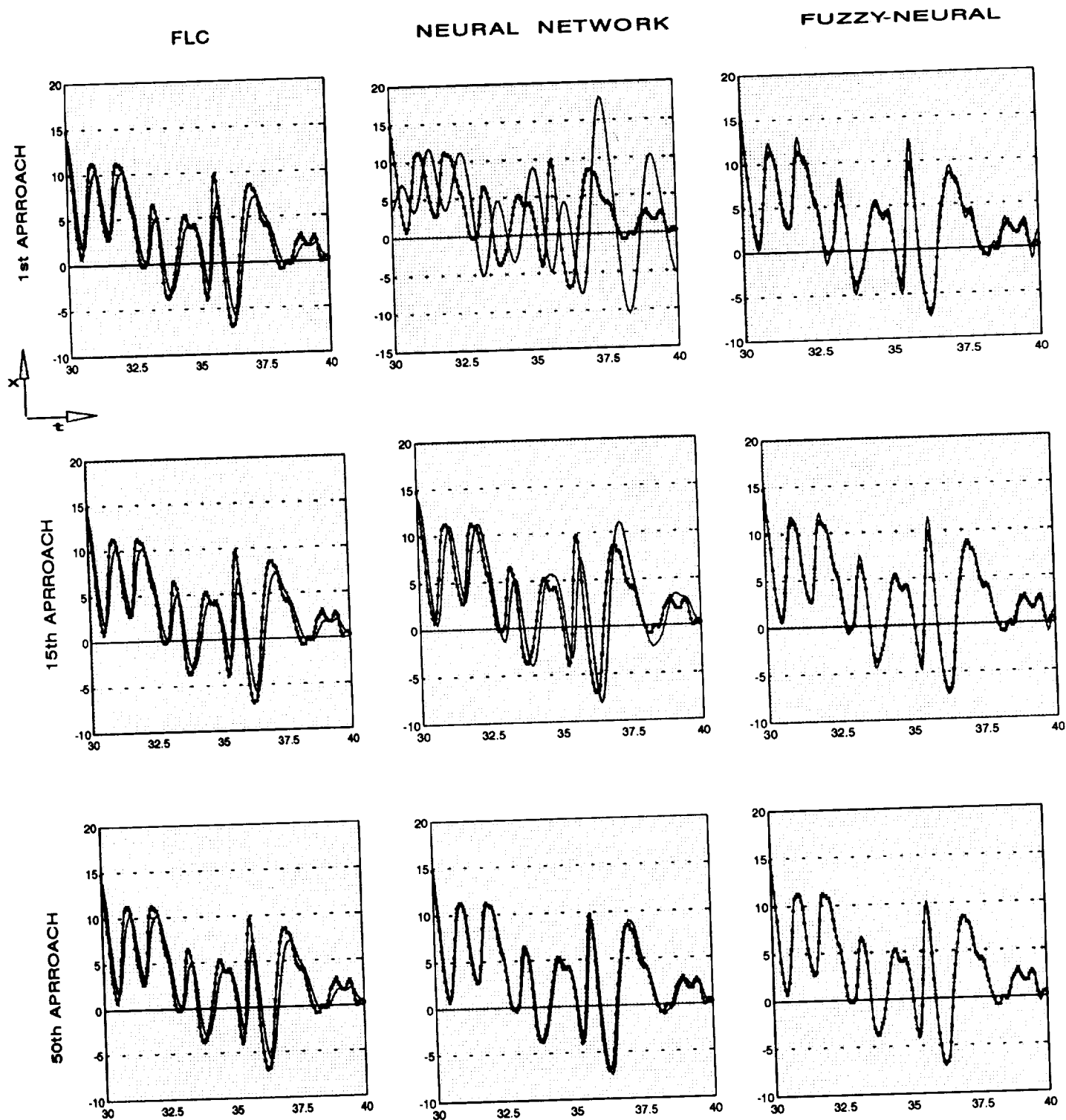


Figure 7 - Comparison of results from FLC, NN, and FNN tracking systems

CONCLUSIONS

A fuzzy-neural network was successfully implemented for automatic camera platform positioning, eliminating the need to manually adjust FLC parameters for optimum performance. The simulations conducted in this study demonstrated that when structured knowledge and learning ability are combined, the result is robust performance that improves over time. The fusion of these two types of machine intelligence represents an important step in the evolution of intelligent control systems.

The FNN may be applied to a wide range of non-linear control systems, particularly where fuzzy control has already proven successful. In any system whose desired behavior can be described through a series of fuzzy rules, those rules can be translated into a network and thereby achieve the ability to adapt. The learning algorithm does not add an excessive amount of computation to an existing FLC, and thus it should be feasible for real-time applications where the system can learn "on the fly". Such experiments are planned for this system in upcoming research.

REFERENCES

1. Kawamura, S., Watanabe, L., Okada, H., and Asakawa K., "A Prototype of Neuro-Fuzzy Cooperation System", *Proceedings of the IEEE International Conference on Fuzzy Systems 1992*, San Diego, CA, 1992
2. Kosko, B., Neural Networks and Fuzzy Systems, Prentice Hall, Englewood Cliffs, NJ, 1991
3. Lea, R., Chowhudry, I., Shehadeh H., and Jani, Y.: "Design and Performance of the Fuzzy Tracking Controller in Software Simulation", *Proceedings of the IEEE International Conference on Fuzzy Systems 1992*, San Diego, CA, 1992
4. Pao, Y., Adaptive Pattern Recognition and Neural Networks, Addison-Wesley Co. Inc., 1989.

54-63

7918

N94-32424

2487

P-10

EMPIRICAL MODELING FOR INTELLIGENT, REAL-TIME MANUFACTURE CONTROL

Xiaoshu Xu
American Welding Institute
10628 Dutchtown Rd.
Knoxville, TN, 37932

ABSTRACT

Artificial neural systems (ANS), also known as neural networks, are an attempt to develop computer systems that emulate the neural reasoning behavior of biological neural systems (e.g. the human brain). As such, they are loosely based on biological neural networks. The ANS consists of a series of nodes (neurons) and weighted connections (axons) that, when presented with a specific input pattern, can associate specific output patterns. It is essentially a highly complex, non-linear, mathematical relationship or transform. These constructs have two significant properties that have proven useful to the authors in signal processing and process modeling: noise tolerance and complex pattern recognition. Specifically, the authors have developed a new network learning algorithm that has resulted in the successful application of ANS's to high speed signal processing and to developing models of highly complex processes. Two of the applications, the Weld Bead Geometry Control System, and Welding Penetration Monitoring System is discussed in the body of this paper.

INTRODUCTION: ARTIFICIAL NEURAL SYSTEMS

Artificial Neural Systems (ANS) are loosely based on biological neural networks offer a computer technology that is a useful tool in process modeling and signal processing. The ANS consists of a series of nodes (neurons) and weighted connections (axons). As with a biological neural network, the assignment of the values of the weights and the size and configuration of the network is the key to a successful net. Unfortunately, we have only begun to understand the inner workings of these constructs. Consequently, relatively crude tools are currently employed to develop working networks.

Typically, the approach to model development is to develop a thorough understanding of the underlying basic scientific or engineering principles of the process. Then, a model is developed that is based on mathematical relationships inherent in the process parameters. The principal drawback with this approach is the time and effort required to develop an understanding of these basic scientific or engineering relationships. Depending on the complexity of the problem, it can take many years and an extensive research program to develop the relationships. Often, instead, a number of simplifying assumptions are made and an approximate model is developed. That approach, while being an expedient method for developing an approximate model that could be useful, provides only a theoretical approximation that may not be valid for the actual problem application.

The artificial neural system, when a network can be found to solve the problem, provides an accurate model of the process or signal. Neural networks are empirical bases systems that, when presented with a specific input pattern, can associate specific output patterns. It is, essentially, a highly complex, non-linear, mathematical relationship or transform. However; it is not necessary for the developer of such a system to understand the basic underlying principles of a process in order to develop a highly accurate ANS based model of the process. Thus, in this way it is quite different from other mathematical modeling approaches.

Basic Principles

The problem of ANS's is to decide how many nodes and connections are needed to model a specific problem, to decide how to configure them, and to decide the specific values of the connection weights and the transfer functions that exist within the network. Figure 1 shows a schematic diagram of a neural network and Figure 2 is a simple representation of the weights, transfer functions, and the mechanisms of network operation. There is no direct known correspondence between the network parameters and its operation and the problem to be modeled by the network. As a consequence, there is currently a lack of good mechanisms which can be used to assign the weights and transfer functions in the network so that it can solve a problem. The methodology of finding a proper net configuration and weights to model a given problem is called the "Learning Algorithm". There are many learning algorithms that have been developed. They all have some advantages and restrictions. In the modeling area, the back-propagation method is the most popular one.

The back propagation method assumes that the search in weight space for an optimum, or near optimum, network configuration can be accomplished as an iterative search using the error gradient, (i.e. slope of the error surface) in L_2 space. That is, a series of moves are accomplished on the multi-dimensional error surface using approximately, the maximum mean squared error gradient direction as the move direction at each iteration. The error, also called the delta, in the network is defined as the mean squared error between the desired output representation and the actual output given the current weight matrix values. By calculating the maximum gradient of the delta for any given training example (set of input and corresponding output patterns), the weights are adjusted so that the net moves along that gradient direction in each presentation of the training example to the network. Using this procedure, the network slowly "learns" to associate all of the training example input patterns with the correct corresponding output patterns by finding a global minimum on the error surface. Since the primary driving force in back-propagation method is the mean squared error - delta, it also called the "delta rule."

This "basic" back propagation learning process has several significant drawbacks. First, the optimum configuration (i.e. number and relative location of hidden representation units or nodes) cannot be pre-determined and, yet, needs to be pre-assigned by using an "educated guess" in order to use this procedure. Since the node configuration can significantly affect the operation of the network, this will at best lead to a long series of re-tries and, at worse, to no useful network at all. Second, this process is very slow and the rate of learning (convergence to near zero error) is set arbitrarily -- traditionally at a value between zero and one. Even though several researches are going on, no currently known method for predetermining the learning rate (gain term) will consistently choose an optimum value, and the optimum value is significantly influenced by the specific problem being presented to the network. Third, it has been shown, that it generally, requires a larger network to "learn" a problem than is required to solve the problem. Based on many study, people generally know that the net could be trim to smaller size. However, there is no known method of reducing the size of the network optimally after training to optimize the net performance. Finally, learning instabilities exist in nearly every problem which will cause the network to stop learning (converging). One of these instability types is known as a local minimum. A local minimum is a depression in the error surface, but one which is not the best minimum or lowest error position. In its search routine, the network algorithm may fall into a local minimum and since the error gradient is toward the local minimum from all directions, may not be able to exit from it.

The Delta Activity Network

A new method was developed by the authors for training neural networks that has been shown to overcome all of the known problems with the back-propagation method, while maintaining the inherent stability and known network development capabilities of that method. This network was developed through the use of a thermodynamic model of the network operation which included both the delta energy and the activity or kinetics of the network. The technique, known as the Delta-Activity Network (D-A Net), has been used on several applications ranging from high speed signal processing to vision systems.

The authors have been studying the learning behaviors of artificial neural networks for years. Based on the thermodynamic and kinetic model of the learning, we discovered that the learning is not only driven by the delta, but also by another important factor - the activity. The activity is a kinetic measurement about how good the neural networks is willing to learn at certain stage of the learning. Sometimes, even the delta is large, but if the system has low activity, it won't learn. This is the case that local minimum or other kinds of learning instabilities accrues. Thus the delta-activity algorithm will watch the activity closely during the learning, maintain a reasonable activity while push the learning rate as high as possible.

The D-A Net has achieved learning rates as high as 1000 times that of the back propagation method while also preventing the network from falling into learning instabilities. Research conducted on this technique has confirmed the existence of at least three types of learning instabilities (local minimum being one of them) and the algorithm can avoid all three instabilities. In addition, using activity as the critical, the system configures itself dynamically during the learning process and so it can produce a near optimum network size for operation, often much smaller than the network needed to "learn" the problem.

This network has been successfully applied to model many real manufacture application problems. These applications are:

- ANS models of weld bead geometry in several different welding processes and joint types
- ANS model to predict weldment mechanical properties for SAW welding
- ANS model of weld metal hot cracking
- ANS model of heat flow for arc weld
- ANS based weld seam tracking system
- ANS based welding penetration monitoring system
- ANS based acoustic emission signal analysis to detect the welding failures
- ANS based underwater acoustical signal detection and classification
- ANS based ultrasonic signal process to detect helicopter tail rotor gearbox fault

In this paper, we will select few examples from above and given more detail discussions.

An ANS Model of Weld Bead Geometry

Background

The mechanical properties, geometry, and appearance of the weld bead are the three major characteristics of the final quality of a weld. They are interactive with each other. The weld bead geometry directly effects the mechanical properties of the weldment. A good weld bead geometry is necessary to insure a good quality weld. The weld bead geometry can be effected by various welding variables. These variables are:

- The process variable, such as the type of welding process, the welding current, voltage, travel speed and wire feed speed.
- The material variables, such as the type and size of base metal and fill metal.
- The joint configuration, such as the type and geometry of welding joint, and the welding position.

To model the bead geometry, an ANS has to perform the transformation from these variables to the bead geometry.

Approach

Four neural networks have been developed to model Gas Tungsten Arc Welding (GTAW), Gas Metal Arc Welding (GMAW), Flux Cored Arc Welding (FCAW), and Submerged Arc Welding (SAW). To simplify the problems, some variables are fixed. The GTAW net is developed based on a stainless steel butt joint weld. The input variables are current, voltage, wire feed speed, travel speed, plate thickness, and joint gap. The FCAW and GMAW nets are based on a plain carbon steel fillet weld.

Several experiments were done to generate the training and test samples. The following table lists the number of training samples used by each net:

Table I
The Number Of Training Samples Used By Each Net.

<u>Net</u>	<u>Number Of Training Samples</u>
GTAW	28
FCAW	34
GMAW	14
SAW	38

Each sample was sectioned and the cross section of the weld bead was measured following the definition to generate the training and test examples.

We have to point out here how few samples the ANS needed to develop a complete model of the weld bead geometry. For instance, in the GTAW case, if we used the Taguchi method, which is known as the best statistical method, we needed at least 150 samples to develop a complete model. Using ANS approach, only 28 samples were needed.

Another important function of the shell is its ability to give the user a suggestion about further experimental points. This function can insure the generation of a complete model with minimum amount of samples.

Results and Discussions

Figure 1 is the configuration of the GTAW net. The net takes 6 inputs and generates 4 outputs. It finally configured itself to have three hidden layers with 7 nodes in each layer.

Figure 2 shows the interface shell of GTAW net. There are two rows of mouse sensitive slide bars on the screen that can be set by the user or observed by the user. These bars can be moved to new values by using the mouse to "slide" the bar. The first row of the slide bars are the outputs of the net, which are the bead width, height, penetration, and the bottom bead width. The second row of slide bars are the input variables, they are the current, voltage, travel speed, wire feed speed, gap, and the plate thickness. At the top left corner is a graphical simulation of the bead cross section. When the net is running in the forward model, the user can slide the input bars, the output bars will change correspondingly and the graphical simulation will animate the bead cross section simultaneously. When the net is running in the inverse model, the user can set the bead geometry by sliding the top row bars and the gap and thickness by sliding the right two bars on the bottom row. The net will find the closest match bead and give the welding parameters which can be read from the bottom row of bars. The whole system runs in real time on a personal computer. Thus, this represents a very powerful planning tool for a welding engineering workstation as well as a complete model for real-time intelligent control.

Figure 2 and 3 show the performance of the GTAW model. In these two figures, the test sample, which the net has not been trained with, is compared with the prediction of the net. A very close match between the predictions of the net model and the actual welds is shown.

Figure 4 shows the GTAW net running in the inverse model. There are two graphical representations of bead cross section on screen. The user specified bead geometry is displayed as the top one. The bottom one is the closes match that the net found.

The GMAW network has similar interface as the GTAW system. However, it also takes two discrete inputs: type of electrode and type of shielding gas. And it also generate two additional discrete outputs: the spade and the silicon slag appearance. FCAW and SAW nets also have a similar interface.

These model networks are the heart of an intelligent control system for welding applications. The objective of an intelligent control system is not to control the primary independent welding parameters (e.g. voltage, current, travel speed, etc.) but to control the final weld quality. That is, control of weld quality parameters such as bead appearance, penetration, amount of spatter, etc. is the goal of an intelligent control system. Sensor data is not particularly useful unless it can be both analyzed and used to control the weld quality parameters. These model networks provide that capability.

The results of these models are being used to develop fuzzy logic based control systems on two projects including development of an intelligent automated welding system for the United States Navy Manufacturing Technology Program called WELDEXCELL.

An ANS Based Welding Penetration Monitoring System

Background

The development of real-time sensing and control techniques for weld penetration has been an active area of research for a number of years. A wide range of techniques have been employed, including pool oscillation measurements, surface infra-red measurements, determinations of the light emitted on the back side of a weld, spectroscopic determination of the presence of tracer elements, and the use of vision techniques to measure weld pool geometry, with the objective of relating to weld penetration.

The success of each of these methods has been limited, and all require the use of specialized transducers. The most successful technology to date appears to be the use of photo detectors to measure the amount of light produced at the back side of the weld, which is then correlated with penetration. Unfortunately, back side access is required, limiting suitability for many applications. Back side monitoring is difficult to make and it is costly.

Approach

Recently, several research works (see ref. [1-3]) have been done to study the weld pool oscillation. The researches show that a full-penetrated weld pool can be modeled as "rubber band" (no rigid base to support it) while a partially penetrated weld pool can be modeled as a half sphere. As a consequence, the pool oscillation pattern will be different in two models. Several methods have been applied to sensor the pool oscillation pattern in real-time, such as the lasers, and trying to relate them with the weld penetration. Unfortunately, the sensors are usually too expensive, and the data from those sensors are too "noisy" to process. These barriers limit the success of those approach.

The authors developed a new approach to detect the pool oscillation pattern and related it to the penetration in real-time. This uses a very in-expensive sensor, the voltage sensor, and the Delta-Activity ANS to process the sensor data. The theory behind this is that when the weld pool oscillates, the voltage between the base metal and welding torch tip will change accordingly. By monitoring the voltage data in a continuous time sequence, the pool oscillation pattern could be detected. However, the voltage data obtained are usually very noisy and contained much more information than just the pool oscillation pattern. Delta-Activity ANS played an important role here to successfully filter out the pool oscillation pattern information from the rest.

The voltage signal is collected at a 1000 Hz sampling rate. A neural network was developed off-line. This neural network takes 100 data points as the inputs. The output of the ANS will indicate that either a full penetration (0) or a lack of penetration (1) is detected. The experimental work used GTAW process for carbon and stainless steel butt joint welding on various plate thickness. Total 25 welding samples were used to train the net.

Result

After training finished, the ANS is installed on a 486 based PC. A GUI is developed to run the penetration monitoring system. If a lack of penetration is detected, the system will turn on a warning light and/or shutdown the welding process. Eventually, combining with the weld bead geometry model, the control loop could be closed and a corresponding welding parameter could be adjusted automatically to insure a full penetration welding. This system is installed in AWI's workshop. Figure 5 is the computer GUI and Figure 6 shows that the system is monitoring a welding.

REFERENCES

- [1]. Y. H. Xiao and G. den Ouden, Weld Pool Oscillation during GTA Welding of Mild Steel. p428-s, Welding Journal, Aug., 1993.
- [2]. Y. M. Zhang, et al. Determining Joint Penetration in GTAW with Vision Sensing of Weld Face Geometry. p463-s, Welding Journal, Oct. 1993.
- [3]. A. A. Shirali and K. C. Mills, The Effect of Welding Parameters on Penetration in GTA Welds. p347-s, Welding Journal, 1993.
- [4]. Pierre Buldi and Kurt Hornik, Neural Network and Principal Component Analysis: Learning from Examples Without Local Minimal, Neural Networks, Vol. 2, pp. 53-58, 1989.
- [5]. J. J. Helferty, J. B. Collins, and M. Kam, A Neuromorphic Learning Strategy for the Control of a One-Legged Hopping Machine, International Joint Conference On Neural Networks, Vol. II, pp. 621, 1989.
- [6]. T. Troudet and W. Merrill, Neuromorphic Learning of Continuous- Valued Mappings in the Presence of Noise: Application to Real-Time Adaptive Control, International Joint Conference On Neural Networks, Vol. II, pp. 621, 1989.
- [7]. E. Barnard and D. Casasent, Image Processing for Image Understanding with Neural Nets, International Joint Conference On Neural Networks, Vol. I, pp. 111-116, 1989.
- [8]. B. R. Kammerer and W. A. Kupper, Design of Hierarchical perceptron Structures and Their Application to the Table of Isolated Word Recognition, International Joint Conference On Neural Networks, Vol. I, pp. 124-150, 1989.
- [9]. A. J. Worth and R. R. Spencer, A neural Network for Tactile Sensing: The Hertzian Contact Problem, International Joint Conference On Neural Networks, Vol. I, pp. 124-150, 1989.
- [10]. R. Fujii, M. F. Tenori, and H. Zhu, Use of Neural Nets in Channel Routing, International Joint Conference On Neural Networks, Vol. I, pp. 321-326, 1989.
- [11]. A. Khotanzad, J.H. Lu, and M. D. Srinath, Target Detection Using a Neural Network Based Passive Sonar System, International Joint Conference On Neural Networks, Vol. I, pp. 335-340, 1989.
- [12]. Richard P. Lippmann, An Introduction to Computing with Neural Nets, IEEE Assp Magazine, April, pp.

4-22, 1987.

- [13]. Back-Propagation -- A generalized delta learning rule, BYTE, pp. 155-192, October, 1987.
- [14]. Stephen Grossberg and Nestor A. Schmajuk, Neural Dynamics of Adaptive Timing and Temporal Discrimination During Associative Learning, Neural Networks, Vol. 2, pp. 79-102, 1989.
- [15] X. Xu, A. Rock, and J. Jones "Neural Network Simulation for Welding Image Understanding" Proceedings, First INNS conference, Minneapolis MN, Aug. 1988.
- [16] X. Xu, J. Jones "A New Computerized Technique for Calculating Ferrite Content", Presented at 1989 AWS 70th annual conference, Washington, D.C. April, 1989.
- [17] A. Rock, X. Xu, and J. Jones "Neural Network Applications in Automated Visual Weld Seam Tracking", Presented at 1989 AWS 70th annual conference, Washington, D.C. April, 1989.
- [18] X. Xu, A. Rock, and J. Jones "Investigation of an Artificial Neural System for a Computerized Welding Vision System" Proceedings, ASM International "TRENDS IN WELDING RESEARCH", Gatlinburg TN, May, 1989.
- [19] X. Xu, A. Rock, and J. Jones, "Use Of An Artificial Neural System For Welding Sensor Signal Processing" 71st American Welding Society Annual Meeting, April 22-27, 1990, Anaheim, CA.
- [20] X. Xu, A. Rock, and J. Jones, "Accelerated Learning Neural Network For Material Processing Sensor Data Analysis" International Conference & Exhibition On Computer Applications To Materials Science And Engineering (CAMSE'90). Aug. 1990, Tokyo, Japan.
- [21] X. Xu and J. Jones, "A NNS based mathematical model of heat flow in PAW and GTAW welding" 72nd American Welding Society Annual Meeting, April, 1991, Detroit.

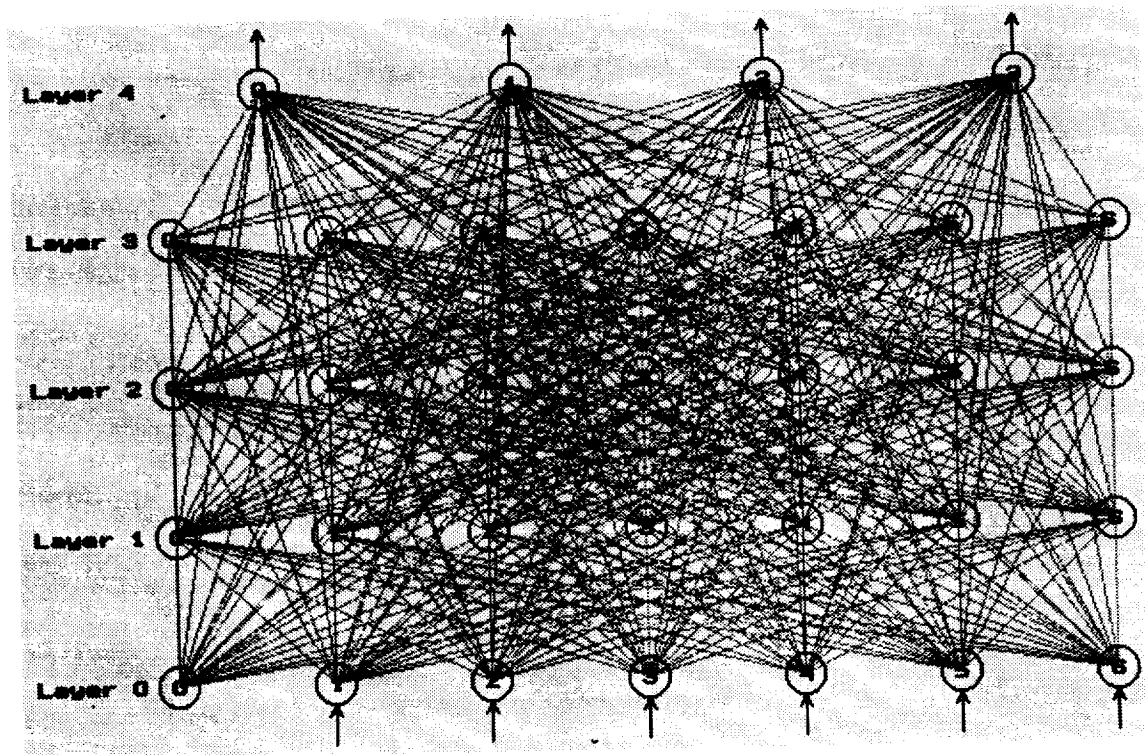


Figure 1. The Configuration of neural network for GTAW modeling

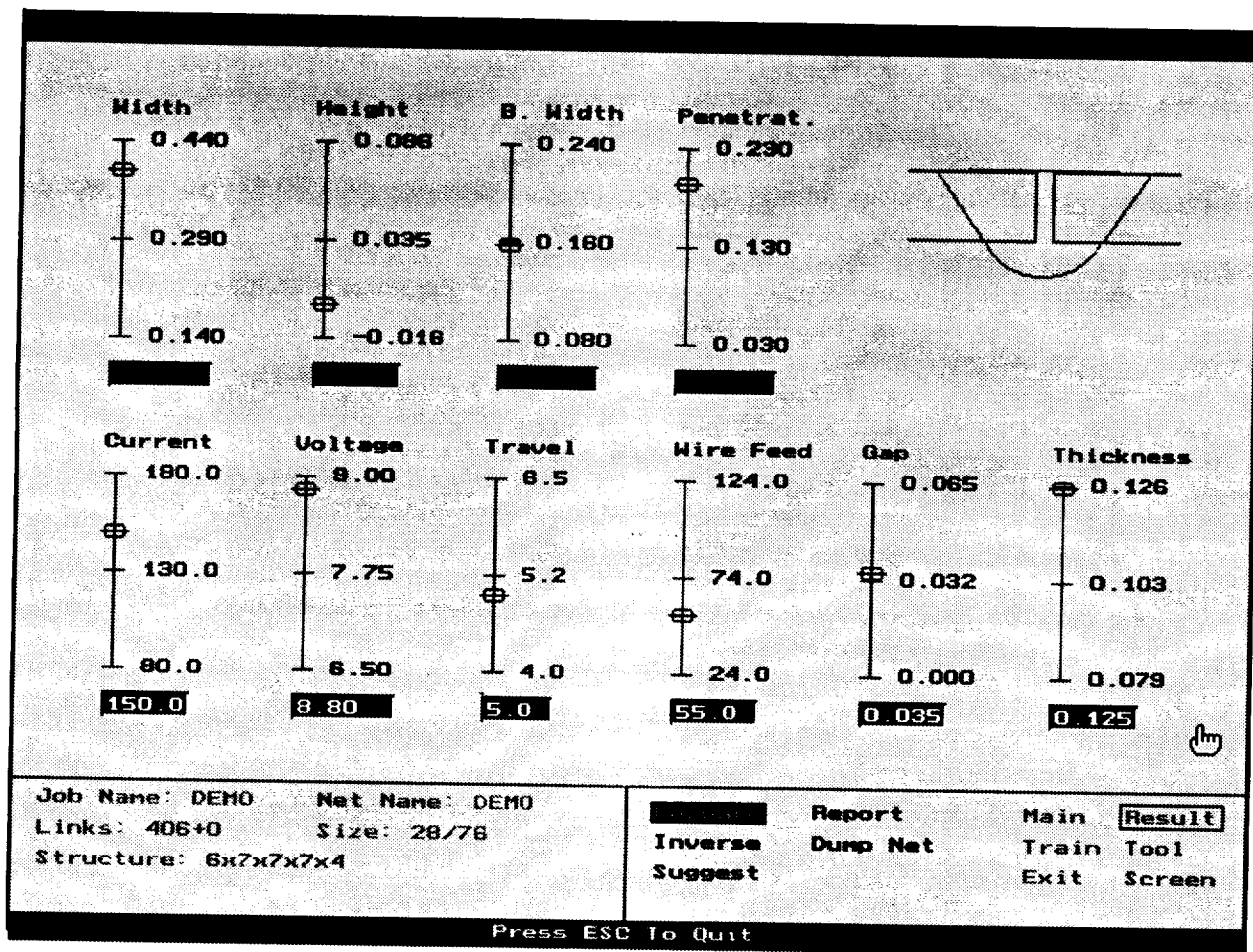


Figure 2. Neural network solution for the GTAW butt weld. Selected parameters of: voltage = 8.8 volts, current = 150.0 amperes, travel speed = 5.0 inches per minute, and wire feed rate = 55 inches per minute. In addition, the weld is being made in material of thickness = 0.125 inches and with a gap of 0.035 inches. The resultant shape parameters and bead graphic are shown in the interface.

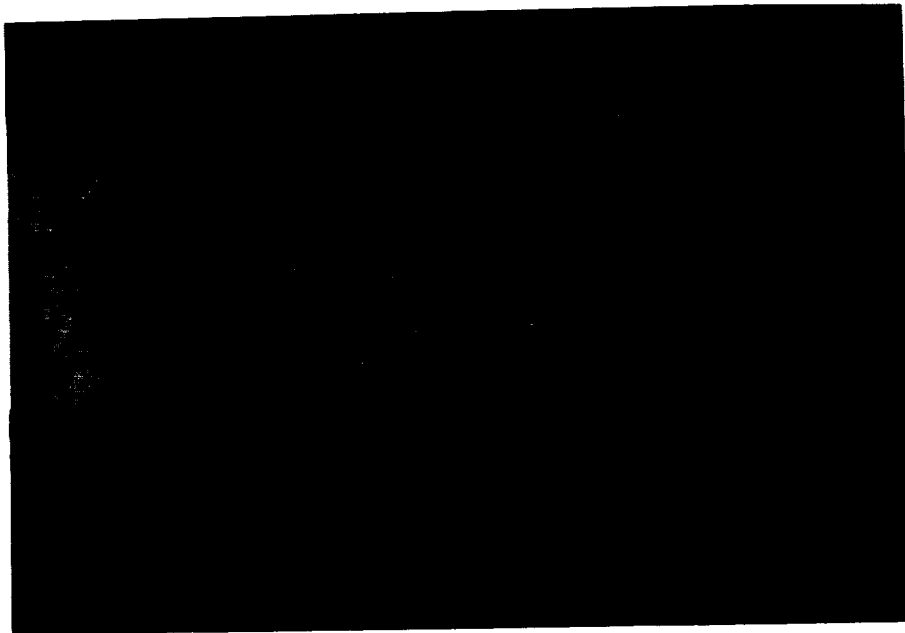


Figure 3. The actual weld bead cross section which was welded using the welding parameters described in the Figure 2.

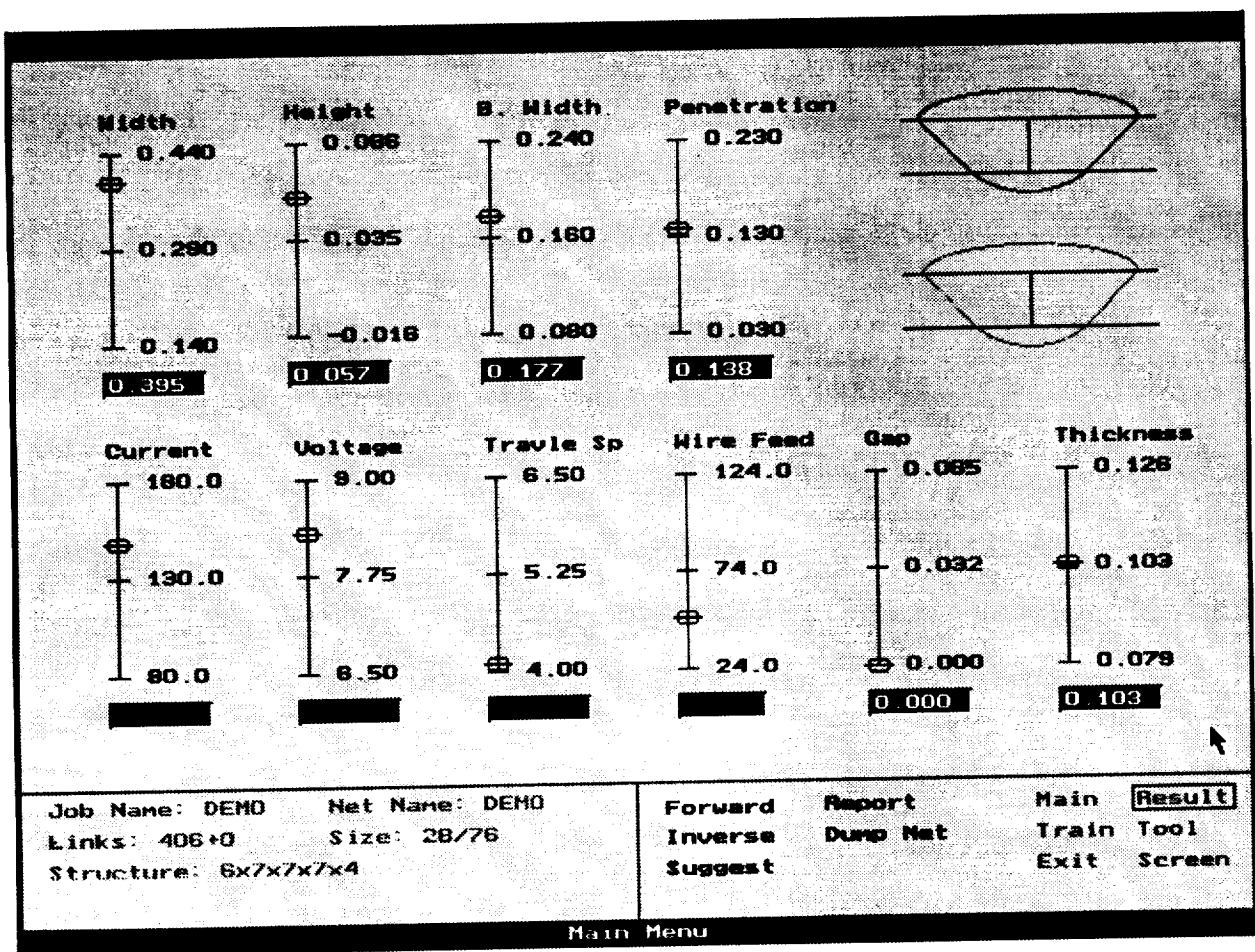


Figure 4. The GTAW network running in inverse model

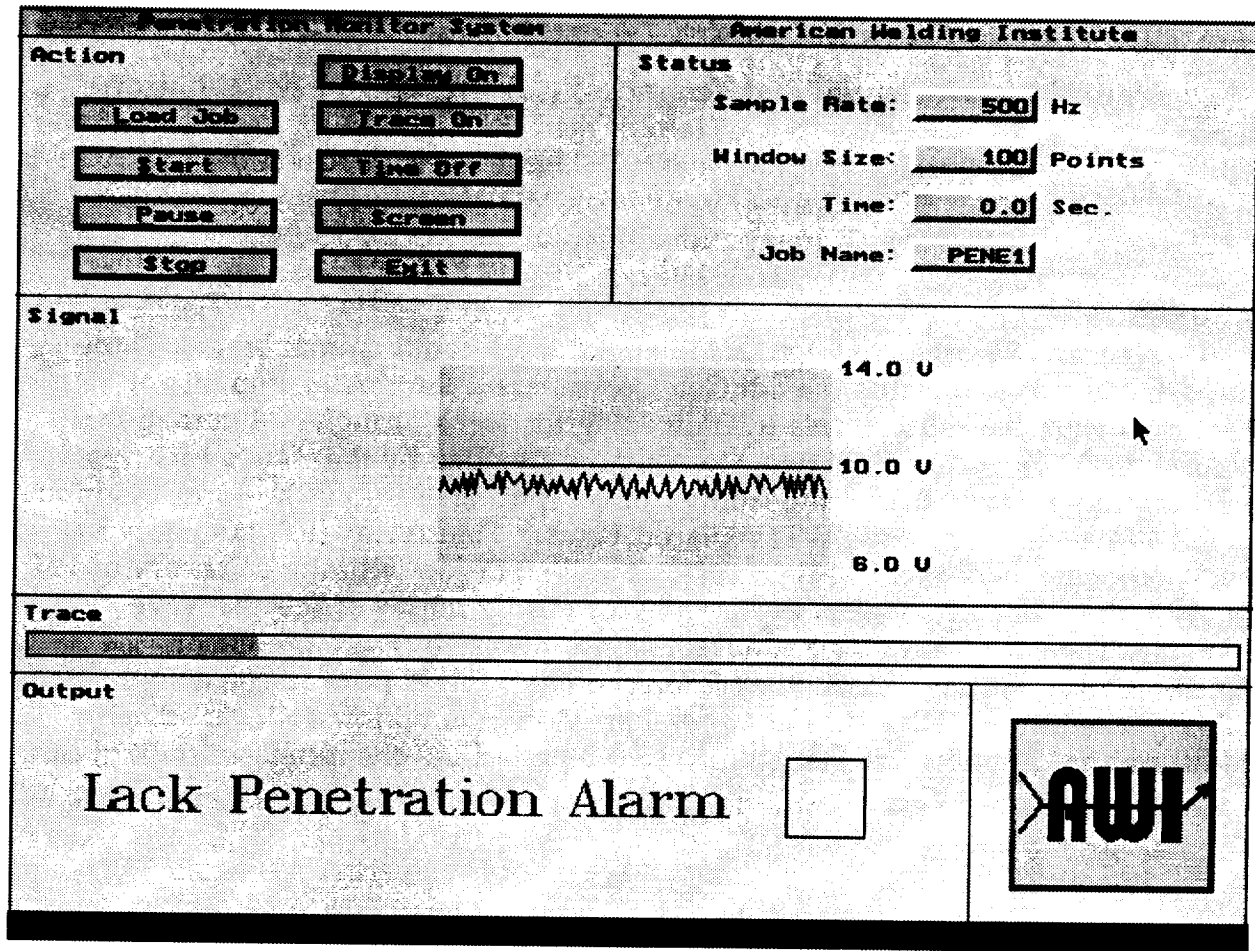


Figure 5. The GUI of real-time welding penetration monitoring system

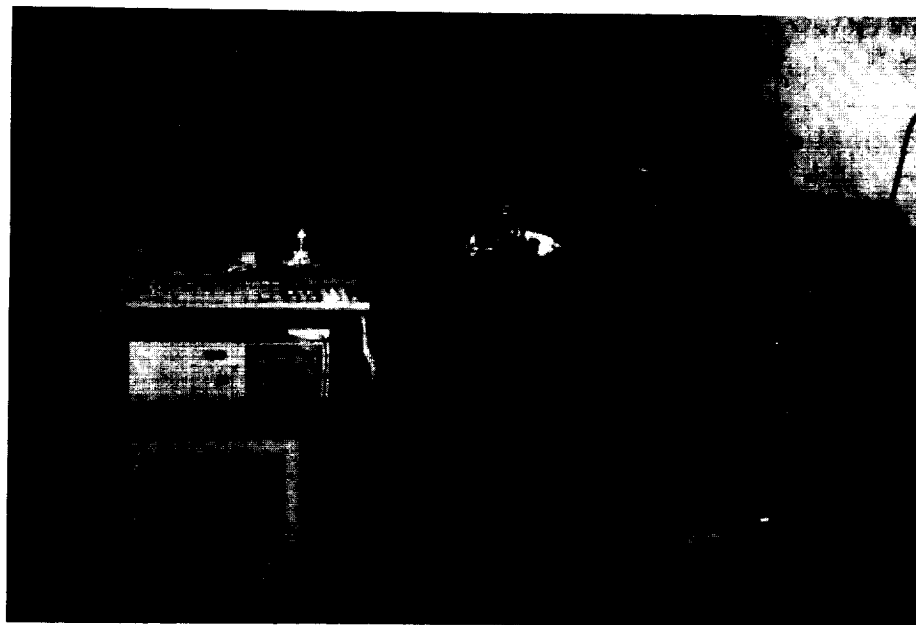


Figure 6. The GTAW Welding Penetration Monitoring System

Neural Network Wavelet Technology, A Frontier of Automation

Harold Szu

Naval Surface Warfare Center Dahlgren Division, Code B44,
Silver Spring/White Oak MD 20903-5640

President, International Neural Network Society
(301) 394-3097; (301) 394-3923(Fax); HSzu@Ulysses.nswc.navy.mil

Abstract:

Neural networks are interdisciplinary studies about animal brains. These have improved AI towards the 6th Gen Computers. Enormous amounts of resources were poured into this R/D awaiting for breakthroughs. International Neural Network Society held two Conferences attended by thousands each year, pushing the ultimate & exciting frontier of computing & info-tech.----our bio-brains,

Wavelet Transforms (WT) replaced Fourier Transforms (FT) favorably in every known Wideband Transient (WT) cases that began with the discovery of WT in 1985--the French geological exploration for oils by means of seismic wave imaging. The list of successful applications has the earth quake prediction, the Radar ID, speech recognitions, stock market forecasts, FBI finger print image compressions, telecommunication ISDN-data compression. More, the billion dollar medical-industrial has applications that still await the perfection--the intelligent heart beat pace-maker, the echoless hearing aids, in vivo constant level drug-dispensor, etc.

1. Introduction

A surging interest in neural nets began in 1980, when J. Hopfield wrote articles in Proc. Nat. Acad. Sci. p.2554 (1992) & p.3088 (1994), although pioneers such as Grossberg, Amari, Widrow, Kohonen, Fukushima, Anderson, Freeman, von der Malsburg, Rumelhart, Werbos, Carpenter, Cooper, have made significant contributions earlier. His model is simple for engineering because of interacting magnets (i.e. neuron points to the north for yes-vote, the south for no-vote) having simple matrix interconnects. Neurons are of McCulloch-Pitts (M-P) threshold logic.

2. Mathematical Foundation of Artificial Neural Networks

M-P Model of the ith neuron (in alphabetic order: u-input & v-output):

$$v_i = \sigma(u_i) = 1/(1 + \exp(-u_i)); \quad \text{Grossberg's: } (d/dt)v_i = c(v_i - \sigma(u_i)) \quad (1)$$

where each voting v_j is weighted by previous memory W_{ij} as the net input u_i :

$$u_i = \sum_j W_{ij} v_j + \theta; \quad \text{Hopfield's: } (d/dt)u_i = b(u_i - \sum_j W_{ij} v_j - \theta); \quad (2)$$

$$W_{ij} = v_i v_j; \quad \text{Hebb's: } (d/dt)W_{ij} = a(W_{ij} - v_i v_j) \quad (3)$$

All lefthand side equations are the **fixed-point solutions of righthand side dynamics**. Such a dynamic, via the synaptic vector outer products: $W_{ij}=v_i v_j$, demonstrated the Zip-Code Sorters, the Bank-Check Readers, and the OCR capabilities [1].

Generalizations in Chaotic-Fuzzy NN chips [21]:

Single neuron processes learnable threshold vector function θ_i .

$$\theta = f(\theta); \quad (d/dt)\theta = e(\theta - f(\theta)) \quad (4)$$

One component of θ has the change detection capability, similar to C. Mead VLSI-analog ANN book (Addison-Wesley 1989). When θ has two first order components giving one second-order equation for an oscillatory behavior similar to firing pulse trains, as shown by Szu et al. Further, if θ has a third component proportional to the need of keeping up the rapid firing, i.e. to the slope of input-output firing rates:

Feigenbaum-Like Chaos Mapping:

$$\theta = \alpha (dv/du) = 4\lambda v(1-v) \quad (5)$$

which gives, explicitly by differentiating Eq(1), the parameter 4λ giving Feigenbaum cascades toward chaos. The model Eq(4) was shown [21] to be a piecewise negative sigmoidal logic, $v_i = \sigma_N(u_i)$ with no delay, which bifurcated for fuzzy uncertainty. I believe that the "Rosetta Stone" of chaos, fuzzy logic, and neural networks has been discovered for which the fuzzy membership function is the triangle envelop of all bifurcation cascade solutions toward the chaos, that covers the full degree from imprecision to precision. The consequence in the chaotic whether prediction that dictates when the cherry in D.C. will be blooming from the imprecise spring season (but a crisp answer yes at the tip of the triangle membership function), to the bifurcation answer in the month of March or April, then further bifurcations into weeks, the days, the hours, etc. Now that a single σ_N neuron can produce such a membership function, of which a hundred thousand of them can learn the collective experience in the chaotic associative memory. So far, we have already demonstrated the habituation and novelty detection effect form such a collection for image processing [21].

3. Mathematics of Wavelet Transforms, and Adaptive WT

Wavelet Transform (WT) is a linear squared-integrable transform in the inner product (Hilbert) space, just like FT. But WT must have a kernel of zero area and vanish outside a finite support. More, WT must generate all localized basis

vectors by the simple affine scale-shift transformation:

$$t \rightarrow t' = (t - b) / a \quad (6)$$

where the dilation $a=2^I$ and the shift $b=I$, are $I=\pm$ integers, rather than by Fourier global harmonics ($2\pi I$ ft). Since WT has the constant fidelity $Q = f_0/\delta f$ of which the ratio between the central frequency f_0 and r.m.s. width of the frequency, WT is analogous to the human inputs. Recently, adaptive WT [2] can determine its own appropriate transform kernel by the input data. A fundamental question is the completeness of the basis vectors. Thus, Adaptive WT is built by Artificial Neural Networks of which each neuron is a daughter. The set of daughter wavelet is

$$\psi_{ab}(t) = \psi(t - b/a) / \sqrt{a}, \quad (7)$$

defined by the mother wavelet $\psi(t)$ satisfying a finite power spectral density (for a derivation of admissible mother, see [3]):

$$\int df |\Psi(f)|^2 / |f| = c_\psi < \infty; \quad (8)$$

where $\Psi(f) = \int dt \exp(2\pi jft) \psi(t)$

The integrand $|\Psi(f)|^2 / |f| \neq \infty$ at $f = 0$, only if $\Psi(f=0) = 0$ gives a zero area condition: $\Psi(f=0) = \int dt \psi(t - b/a) = 0$. The integrand is identical by changing $f \rightarrow f' = af$. All daughters have identical inner product to the mother: $(\psi_{ab}(t), \psi_{ab}(t)) = (\psi(t'), \psi(t'))$.

According to the best possible classification or the most efficient representation criteria, the adaptive WT yields a composed mother from a set of admissible mothers: $\psi^{(n)}(t)$, Eq(3). Will the super-mother $h(t)$ be admissible Eq(8)?

$$\text{Min. } \langle |s(t) - \sum_a \sum_b (s(t), h(t-b/a)) h(t-b/a)|^2 \rangle \quad (9)$$

where $h(t) = \sum_n w_n \psi^{(n)}(t)$.

Prove of Szu's Theorem of Adaptive Wavelet Transform with Admissible Super-Mother Condition:

Since a square modulus is always real positive: $0 \leq |\Psi^{(m)} - \Psi^{(n)}|^2$, then, by expanding the quadratic expression, the Schwartz inequality holds:

$$\Psi^{(m)} \Psi^{*(n)} + \Psi^{*(m)} \Psi^{(n)} \leq |\Psi^{(m)}|^2 + |\Psi^{(n)}|^2$$

Use is made of Eq(4) in FT domain, we have the admissible super-mother condition:

$$\begin{aligned} \int df |H(f)|^2/|f| &= \sum_m d_m \sum_n d_n \int df \Psi^{(m)} \Psi^{*(n)}/|f| \\ &\leq (1/2) \sum_m \sum_n w_m w_n \left\{ \int df |\Psi^{(m)}(f)|^2/|f| + \int df |\Psi^{(n)}(f)|^2/|f| \right\} \leq \infty, \end{aligned} \quad (10)$$

which is bounded. Thus, the super-mother is likewise bounded, i.e. admissible.

Examples:

For the NASA mission of downward looking surveillance, such as the Earth Observation System (EOS), and DoD Mission Planning Precision Striking Initiative, we requires a 2-D imagery sequence in time to be analyzed by 2-D wavelet adaptively. In order to illustrate the AW T, the simplest possible case is a hybrid of the human vision system response function called the Difference of Gaussian (DOG) model and the complex Gabor-Morlet wavelet in 1-D

$$h^{(n)}(t) = \exp(-t^2/a_n) (t^2 - b_n^2) \exp(2\pi j c_n t). \quad (11)$$

where a_n, b_n, c_n are real constants to be determined by ANN, that could be different for different mothers.

Another interesting example is an envelope soliton for the nonlinear ocean dynamics,

$$h^{(n)}(t) = \text{sech}^2(t-b/a) \exp(2\pi j c_n t) \quad (12)$$

which solves the Korteweg-DeVries equation derived from the Navier-Stokes hydrodynamics for a surface wave. Because the soliton is the exact solution of the nonlinear dynamics, this example indicates the potential that WT has---the freedom "to pay the nonlinear price first, and then enjoy the linear superposition" later. On the contrary, FT must be deferred the nonlinearity nature by first apply the linear FT to the nonlinear problem. Then, the challenge of the truncation of the NL mode-mode coupling equations remains, which may be solved only in the limit of weak nonlinearity.

References:

1. K. Schiff and H. Szu, "Gram-Schmidt orthogonalization Neural Networks for Optical Character recognition," J. Neural Network Computing, Vol. 1, No. 3, pp.5-13, 1990 (Auerbach Publ. N. Y.)

2. H.H. Szu, B. Telfer, S. Kadambe, "Neural Network Adaptive Wavelets for Signal Representation and Classification," *Opt. Eng.* Vol.31, pp.1907-1916, Sept. 1992.
3. H.H. Szu, Y. Sheng, J. Chen, "The Wavelet Transform as a Bank of Matched Filters," *Appl. Opt.* Vol. 31, pp.3267-3277, 1992.
4. Y. Sheng, D. Roberge, H. Szu, "Optical Wavelet Matched Filters for Shift-Invariant Pattern Recognition," *Opt. Lett.* Vol.18, No. 4, pp. 299-301, Feb 15, 1993.
5. H. Szu, X-Y Yang, B. A. Telfer, Y. Sheng, "Neural Network and Wavelet Transform for Scale-Invariant Data Processing," *Physics Review E*, Vol. 48, No.2, pp. 1497-1501, August 1993.
6. B. Telfer, H. Szu, R. Kiang, "Classifying Multispectral Data by Neural Networks," *Telematics & Informatics*, Vol. 10, No.3, pp. 209-222, 1993.
7. H. H. Szu, B. Telfer, A. Lohmann, "Causal Analytical Wavelet Transform," *Opt. Eng.* Vol. 31, pp. 1825-1829, Sept. 1992.
8. B. Telfer, H. Szu, "New Wavelet Transform Normalization to remove Frequency Bias," *Opt. Eng.* Vol.31, pp.1830-1834, Sept. 1992.
9. H.J. Caulfield, H. Szu, "Parallel Discrete and Continuous Wavelet transforms," *Opt. Eng.* Vol. 31, pp. 1835-11939, Sept.1992.
10. Y. Sheng, D. Roberge, H. Szu, "Optical Wavelet Transform," *Opt. Eng.* Vol. 31, pp.1840-1845, Sept. 1992.
11. X. Yang, H. Szu, Y. Sheng, H.J. Caulfield, "Optical Haar Wavelet transforms of Binary Images," *Opt. Eng.* Vol. 31, pp. 1846-1851, Sept. 1992.
12. G. Rogers, J. Solka, C. Priebe, H. Szu, "Optoelectronic Computation of Wavelet like-based Features," *Opt. Eng.* Vol. 31, pp. 1886-1893, Sept. 1992.
13. T.K. Oh, N. Caviris, Y. Li, H. Szu, "Texture Analysis by Space-Filling Curves and 1-D Haar Wavelets," *S. Phuvan*, *Opt. Eng.* Vol. 31, pp. 1899-1906, Sept. 1992.
14. H. Szu & B. Telfer, In: "MIT Handbook on Neural Networks", Chapter on Dynamic Wavelet Transform
15. H. Szu, "Why the Soliton Wavelet Transform is useful for Nonlinear Dynamic Phenomena, *SPIE Proceedings*, Vol. 1705, pp. 280-288, 1992.
16. M. Bodruzzaman, X. Li, K. Kuah, H. Szu, B. Telfer, "Speaker recognition Using Neural Network and Adaptive Wavelet Transform," *Proceedings of SPIE Vol. 1961*, Orlando April, 1993
17. B. Telfer, H. Szu, A. Dubey, N. Witherspoon, "Detecting Blobs in Multispectral Electro-Optical Imagery Using Wavelet techniques," *Proc. SPIE Vol. 1961*, Orlando April 1993.

18. S. Kadambe, B. Telfer, H. Szu, "Representation and Classification of Unvoiced Sounds using Adaptive Wavelets," Proc. SPIE Vol. 1961, April Orlando 1993.
19. H. Szu, B. Telfer, R. Kiang, "Neural Network Classification of Multi-spectral Data with Contextual Information," WCNN Portland July 1993.
20. B. Telfer, H. Szu, R. Kiang, "Classifying Multispectral Data by Neural Networks," NASA AI Goddard Conference, May 12, 1993.
21. H. Szu, L. Zadeh, C. Hsu, J. Dewitte, G. Moon, D. Gobovic, M. Zaghloul, "Chaotic Neurochips for Fuzzy Computing", Proc. of SPIE, Vol. 2037, S.D., July, 1993.
22. H. Szu, G. Rogers, "Single Neuron Chaos," IJCNN -92 Baltimore, V .III, pp103-108.
23. H. Szu, G. Rogers, "Generalized McCullouch-PittsNeuron Model with Threshold Dynamics," IJCNN-92, Vol. III, pp505-510. Baltimore, Vol. III, pp535-540 (June 7-11,92)
24. H. Szu, B. Telfer, G Rogers, Kyoung Lee, Gyu Moon, "Chaotic Systems," Proc. IEEE, V. 75, no.8, M.Zaghloul, M. Loew," Collective Chaos in Neural Networks," Int'l Joint Conf. Neural Networks, IJCNN-92 Beijing China, Nov 1-6, 1992.
25. H. Szu, B. Telfer, G. Rogers, D. Gobovic, C. Hsu, M. Zaghloul, W. Freeman, "Spatiotemporal Information Processing in Chaotic Neural Networks -- Electric Implementation," World Congress of Neural Networks, WCNN-93, Portland OR, July 12-16, 1993.

NEW APPROACHES FOR REAL TIME DECISION SUPPORT SYSTEMS

D. Charles Hair and Kent Pickslay
NCCOSC RDT&E Division, Code 444
53140 Gatchell Road, Room 421A
San Diego, CA 92152-7420
email: hair@popeye.nosc.mil
phone: 619-553-5302
fax: 619-553-4149

ABSTRACT

NCCOSC RDT&E Division (NRaD) is conducting research into ways of improving decision support systems (DSS) that are used in tactical Navy decision making situations. The research has focused on the incorporation of findings about naturalistic decision-making processes into the design of the DSS. As part of that research, two computer tools have been developed that model the two primary naturalistic decision-making strategies used by Navy experts in tactical settings. Current work is exploring how best to incorporate the information produced by those tools into an existing simulation of current Navy decision support systems. This work has implications for any applications involving the need to make decisions under time pressure, based on incomplete or ambiguous data.

BACKGROUND

Our research at NRaD is part of the TADMUS (Tactical Decision Making Under Stress) project, funded by the Office of Naval Research. That project is generally involved with looking into new ways to enhance the decision making of Navy personnel in tactical command and control situations.

The particular focus of TADMUS is on the area of anti-air warfare (AAW), and involves situations in which shipboard commanders must make decisions about the nature and intent of aircraft that are in the vicinity of the ship. Because these situations can involve jet aircraft armed with missiles, decisions must sometimes be made in seconds even though the available information can be incomplete or ambiguous. The decisions are made by six person teams, where the ultimate responsibility for making decisions lies with the ship's commander (CO) who is closely aided by a tactical action officer (TAO).

The TADMUS research has taken two primary directions. Part of the work is involved with looking at new training approaches for the decision teams. The other primary effort involves the investigation of new approaches in the area of decision support systems. We are involved with the DSS area at NRaD.

A large part of the early DSS related work under TADMUS involved the investigation of the decision-making strategies used by experienced Navy personnel in actual tactical situations. That research took the approach of looking into naturalistic decision making. Naturalistic decision making emphasizes gathering data about how experienced decision makers make their decisions in real world settings. This approach is to be distinguished from the more artificial approaches often used in decision research, where inexperienced subjects are tested in doing unfamiliar tasks in artificial settings. One of the general ideas emerging from studies of naturalistic decision making is that it appears that experienced human decision makers are much better at making good decisions than is often suggested by more traditional research approaches.

The early TADMUS research identified two primary naturalistic decision-making strategies being used by experienced Navy personnel [5]. In most situations, those decision makers use a strategy referred to as recognition-primed decision making (RPD). This strategy relies on the use of prior experience to suggest how prototypical patterns of data may be sufficiently close to currently observed data to guide decisions about the current data [6]. When prior experience proves insufficient to guide current decisions, a second strategy comes into play using explanation-based decision making, or story generation. This strategy involves the construction of a few alternative explanations that account for the current data, followed by an evaluation of which explanation is the most plausible [8]. In both the RPD and story generation situations, selection of a course of action is expected to be automatically generated as a consequence of the situation assessment.

Two tools have been developed under TADMUS to model those strategies. An RPD template based tool takes the approach of matching current data items against a set of stored templates [7]. Those templates represent a set of typical scenarios that can be expected to occur in given tactical settings. The tool estimates the degree of fit between actual current data and the stored templates, and brings templates to the user's attention when there is a sufficiently good fit. A second tool called SABER (Situation Assessment By Explanation based Reasoning)

constructs alternative explanations that suggest how all current data can be accounted for in reaching distinct different conclusions [2,3]. SABER makes use of a few heuristics to evaluate which of the alternatives is most plausible.

Also in connection with the TADMUS DSS work, a simulation facility has been set up that models existing Navy shipboard systems. A number of experiments have been run at that facility to develop empirical evidence about what kinds of problems arise [4]. A new set of experiments will begin shortly to compare performances with and without the use of the new tools. All experiments are done with experienced Navy personnel.

NEW IDEAS IN DECISION SUPPORT

New Information Products

One of the key hypotheses relied on in developing the two decision support tools is that the use of models of people's cognitive strategies will promote the ability of the tools to enhance people's performance with those strategies. At a general level, the idea is to match the internal models of the tools to the naturalistic decision-making strategies of the users, as a means of achieving a cognitive fit between the way both the tool and the user approach the problem. That fit is expected to lead to improved performance by facilitating a more rapid and accurate use of the strategies. The desirability of achieving a cognitive fit has been emphasized in recent work in the human-computer interface field dealing particularly with tools intended to enhance human performance [10].

Development of these tools has led to an ability to produce new kinds of information products for the decision makers. Some of the key new information products available from the RPD tool are: scenario events shown in a timeline, predicted events, and suggested responses. These products arise directly out of the tool's use of timeline based templates representing event scenarios.

The more important new products produced by the SABER tool are: an ability to group data according to the hypotheses each piece of data directly supports, the related ability to indicate where important pieces of data are missing, and a review of possible ways to account for contradictory pieces of data. This kind of information arises from the tool's ability to construct alternative explanatory structures to account for arbitrary sets of data.

Some of these new information products are illustrated in Figures 1 and 2, where virtually all of the parts of those figures represent either completely new kinds of information or a new way of presenting the information, as compared to existing Navy decision support systems. These figures show information related to a given aircraft track, where the tracks are typically shown as symbols on a geographic display. Figure 1 illustrates an early proposed interface for the new tools. There are nine separate windows illustrated which in counterclockwise order show the following: (1) a three dimensional display of the track history for a selected track; (2) a list of events of interest for the track, keyed into the 3D display; (3) a list of other information related to the track; (4) a bar graph indicating the most likely hypothesis as to the nature of the track; (5) a template presentation showing the key elements of the hostile intent hypothesis; (6) a geographical display showing various tracks, with the currently selected track highlighted; (8) a list of suggested responses with regard to the track; and, (9) a more general list of alerts, which bring items of possible interest to the user's attention.

Figure 2 illustrates a way of presenting some of the information available through the SABER tool. The window to the right shows the grouping of data items according to four high level possible conclusions. Also shown is the ability to indicate important pieces of data that may be missing, where in that window such items are greyed out. The bottom part of the window to the left shows where assumptions can be used to explain away some pieces of data. The upper portion of that window represents a way of indicating which of the four conclusions is the most plausible according to the SABER tool.

Building the Interface

A major goal in designing the interface is to try to satisfy the partially conflicting goals of providing new kinds of information, while at the same time minimizing the time required for users to assimilate the available information. The need for minimizing user interaction time results directly from the real time nature of the tactical situation. As a result of these goals, we are directing part of the design effort at determining what new information is most critical to the decision maker and at finding ways to compress the presentation of that information.

A related problem lies in presenting the information produced by the system in a way that does not suggest to the user that the actual decisions are being made by the system. It is important to note that since these tactical situations necessarily involve incomplete and ambiguous information, it is not possible for any computer tool to generate completely reliable conclusions. In that kind of situation it seems preferable to leave the ultimate decision up to the user, particularly where the decisions can have life or death consequences. One of the principal questions that needs to be resolved in this connection is whether or not the system should indicate to the user the system's own evaluation of the situation.

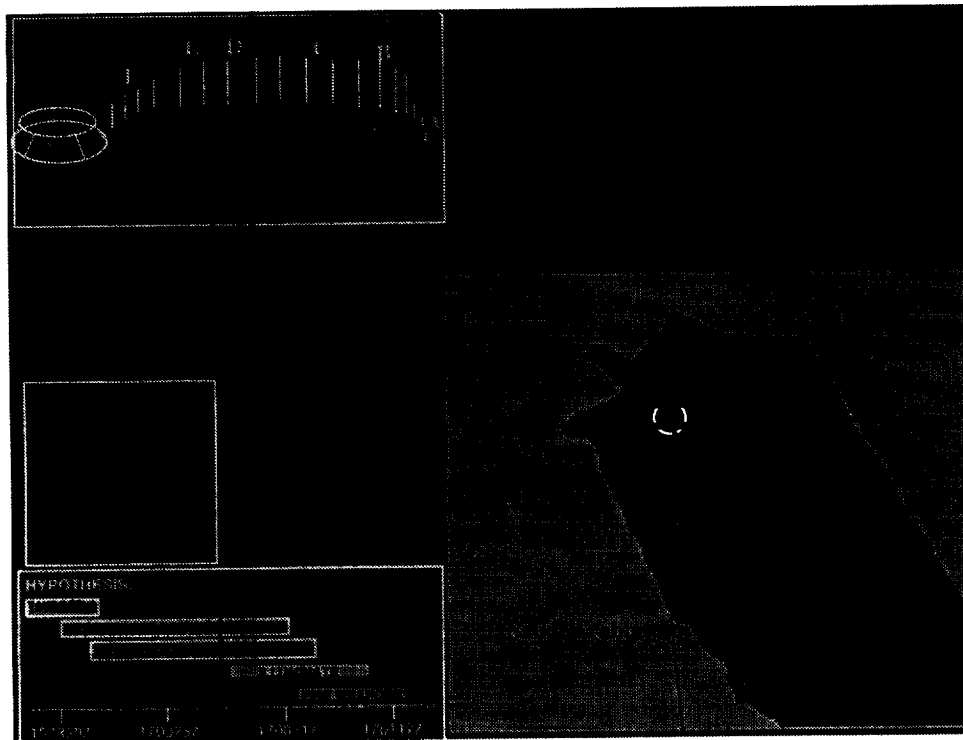


Figure 1. General decision support system interface.

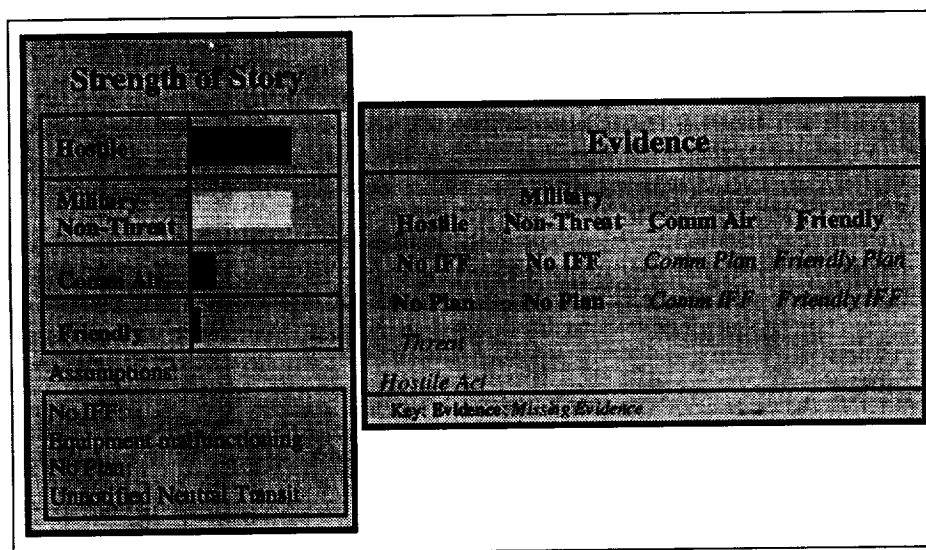


Figure 2. SABER related information products.

In arriving at some final design ideas, we have radically changed our approach to integrating the new information products into the existing DSS framework. The RPD and SABER tools were originally developed as highly interactive, standalone tools. At that time it was expected that the tools would be incorporated into the overall DSS system through separate CRT displays. However, it became clear from both the real time nature of the problem and the identity of the users that that approach was not viable. In fact, it appeared that if that approach were adhered to we would have reached the situation of needing to add a new person to the decision team who could mediate between the two tools and the other team members.

The real time constraint leads to a general need to avoid the need for user interactions with the system. In addition, as the TADMUS empirical work progressed, the decision was made to focus on the two members of the decision making team who have the ultimate decision making responsibility: the CO and TAO. Those individuals

are officers who typically do not take a very active role in interacting directly with the DSS, making it even more desirable not to introduce new interactive features.

The net result was to decide on a single CRT display, that would combine the most critical information products of the RPD and SABER tools. The general idea of the intended product is illustrated in Figure 1. It should be noted that the actual display will be in color rather than the black and white shown in Figure 1. The main point here is that Figure 1 shows that a variety of information can be incorporated in a display on a single monitor.

Also of importance in Figure 1 is the point that the only kind of user interaction called for is to click on a track of interest in the geographical display window. Related information of interest is then automatically changed in the other windows. This action is one that is already required on existing DSS's. It should further be pointed out that the symbology used in the geographical display window shown in Figure 1 is unchanged from standard Navy usage, except that color has been added to our product.

In determining how best to fit together the RPD and SABER tools we have developed the view that the tools fit together in a two level approach. At the more specific level, the RPD tool is able to determine how well current data fits into one or more predefined template representations of typical scenarios. Where there is a good fit, the RPD tool can bring templates to the user's attention along with accompanying information products. The SABER tool is seen as performing a more general role that is useful when current data may not fit any of the predefined templates well. Viewed at this level, the role of SABER is seen as supplying tentative, alternative ways of explaining the data in terms of a limited number of high level possible conclusions.

Critics

In part, our interface design efforts are now being guided by principles related to the use of critics in user interfaces. Critics are devices that have received increasing interest in the user interface community in recent years [1,9]. Critics have been implemented in a number of forms, but the essential idea is to have a built-in mechanism for making suggestions to the user about how to proceed with a given task. This kind of mechanism is particularly useful where the computer and user are thought of as working together to cooperatively solve a problem of some sort. Critics have been most often used in cooperative design settings.

That approach fits in well with what we want to accomplish with the TADMUS DSS research. The goal is to achieve a situation in which decisions are made partly as a result of cooperative problem-solving. What is then needed is to identify which activities are done best by computers and which by people in this setting, with the view that by integrating those activities the overall cooperative performance will exceed what either the computer or the person could do alone.

One way of utilizing the critic concept would be to build a model of the user into the interface such that the system could calculate from the user's actions what the user was trying to do at any given time. Based on that kind of knowledge the critic could automatically supply new information as it reasoned that the user needed it. Although we believe that approach is worth further study, at the present time we are not pursuing it. The problems are that the approach does not appear to be appropriate for real time applications, and in our situations it is likely that users would object to having the system decide for them when to change displays and what to display.

Instead, we have adopted the approach of trying to critique the user's decision-making processes in a way that is purely passive and transparent to the user. Essentially, we try to emphasize through our display that there are always alternative ways of explaining a given set of data. While doing that we show some possible ways in which the existing data fits together, and indicate where the fact that some pieces of data are missing may be important. This approach is illustrated in Figure 2 where the user is shown how a current set of data relates to the top level conclusions about whether an aircraft may be friendly or not. That figure indicates that the interface can compactly show alternative ways of accounting for data as long as there are a limited number of top level conclusions.

POSSIBLE COMMERCIAL APPLICATIONS

There are applications for this research in any setting involving the need for real time decision making. Some examples are: some kinds of medical settings, weather forecasting, and possibly air traffic control. These are situations in which decisions must sometimes be made quickly, and may need to be based on incomplete or ambiguous data.

The specific RPD and SABER tools developed for the TADMUS project are not likely to be of value in other settings, since they have both been specially-tailored for the AAW command and control setting. However, it is reasonable to assume that the decision-making strategies modeled through those tools are the same strategies used by other decision makers in stressful, real time decision making areas. Where those strategies are used, it should be straightforward to create tools that use the basic underlying approaches of the two tools. The part that would not be so easy, would be tailoring the interface to meet the needs of new domains.

There are a few key lessons learned in our project that should be kept in mind in developing new special purpose tools based on these strategies. The most important lesson is that there are essentially two approaches to making new tools like this available to users. One way is to build in a lot of functionality along with interactive features, with the assumption that either the user will spend a significant amount of time learning to use the tools or that a new human assistant will be used to mediate between the user and the tool. The other option is to try to simplify the interface as much as possible, and require as little interaction as possible. We have chosen the second approach at NRaD. A related lesson is that it is vitally important to identify the ultimate intended user as early as possible so that realistic approaches can be taken with regard to the use of features that require extensive interaction.

CONCLUSION

Initial research done under the TADMUS project has produced a theory about the strategies used by expert Navy decision makers in tactical command and control situations. Two computer tools have been developed that are tied into that theory: the SABER tool and an RPD tool. Work is now being done to determine how best to present the information produced by those tools to users. The expectation is that such information will lead to improved decision making.

The principles being developed through this work have application in any areas involving the need for real time decision support systems. The most obvious area of applicability is in medicine where emergency situations can raise problems with time constraints, incomplete data, and ambiguous data. Those are exactly the kind of problems that are the focus of the TADMUS efforts.

Our design process has suggested a few design principles of particular interest in these real time settings. At a general level it can be noted that simply using models of user cognitive strategies may not directly suggest good ways to construct the interface. We believe that the models do lead to producing the kind of information that is needed, but that determining how to present the information remains a hard problem. More specifically, we have found that although direct manipulation is normally viewed as a desirable approach for letting users manipulate data, in this real time setting it is desirable to minimize such manipulations in order to minimize the time the user must spend dealing with the interface. In addition, we are using graphical data presentations where possible, but not to the extent that users are required to learn an extensive new set of symbols.

At a general level we have focused our design efforts on the idea that the RPD and SABER tools fit naturally together through a two level model of the kinds of information that need to be presented to users. The RPD tool can supply information related to specific, detailed representations of scenarios, while the SABER tool relates data to more general possible conclusions. In addition, we are exploiting the idea that our overall interface can be viewed as a type of critic.

REFERENCES

1. Fischer, G., Lemke, A.C., Mastaglio, T., and Morch, A. The role of critiquing in cooperative problem solving. *ACM Transactions on Information Systems*, Vol. 9, No. 2, pp. 123-151, April, 1991.
2. Hair, D.C., Pickslay, K. & Chow, S. Explanation-based decision support in real time situations. In *Proceedings of 4th International Conference on Tools With Artificial Intelligence*, pp. 22-25, Arlington, Virginia, 1992.
3. Hair, D.C. and Pickslay, K. Development of computer support for naturalistic decision-making. Technical Report 1558, Naval Command, Control and Ocean Surveillance Center, RDT&E Division, December, 1992.
4. Hutchins, S.G. and Kowalski, J.T. Tactical decision making under stress: Preliminary results and lessons learned. To appear in *Proceedings of the 1993 Symposium on Command and Control Research*, Washington, D.C., June, 1993.
5. Kaempf, G.L., Miller, T.E., & Wolf, S. Decision requirements for the design of human-computer interfaces. In *Proceedings of 10th Annual Conference on Command and Control Decision Aids*, Washington, D.C., June, 1993.
6. Klein, G.A., 1989. Recognition-primed decisions. *Advances in Man Machine Systems Research* 5, pp. 47-92, Rouse, W.R. (ed.), J.A. Press.
7. Noble, D. and Flynn, W. The TADMUS "RPD tool". In *Proceedings of 10th Annual Conference on Command and Control Decision Aids*, Washington, D.C., June, 1993.

8. Pennington, N. and Hastie, R. Explanation-based decision making: The effects of memory structure on judgement. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, pp. 521-533, 1988.
9. Rettig, M. Cooperative software. *CACM*, Vol. 36, No. 4, pp. 23-28, April, 1993.
10. Smith, J.B. and Lansman, M. Designing theory-based systems: A case study. In *Proceedings of CHI'92 Conference Proceedings*, pp. 479-488, Monterey, California, May, 1992.

7921
27-51
2490
p-a
KNOWLEDGE-BASED COMMODITY DISTRIBUTION PLANNING

Dr. Victor Saks
Carnegie Group, Inc.
Pittsburgh, PA 15222

Ivan Johnson
Carnegie Group, Inc.
Pittsburgh, PA 15222

ABSTRACT

This paper presents an overview of a Decision Support System (DSS) that incorporates Knowledge-Based (KB) and commercial off the shelf (COTS) technology components. The Knowledge-Based Logistics Planning Shell (KBLPS) is a state-of-the-art DSS with an interactive map-oriented graphics user interface and powerful underlying planning algorithms. KBLPS has been designed and implemented to support skilled Army logisticians to prepare and evaluate logistics plans rapidly, in order to support corps-level battle scenarios. KBLPS represents a substantial advance in graphical interactive planning tools, with the inclusion of intelligent planning algorithms that provide a powerful adjunct to the planning skills of commodity distribution planners.

INTRODUCTION

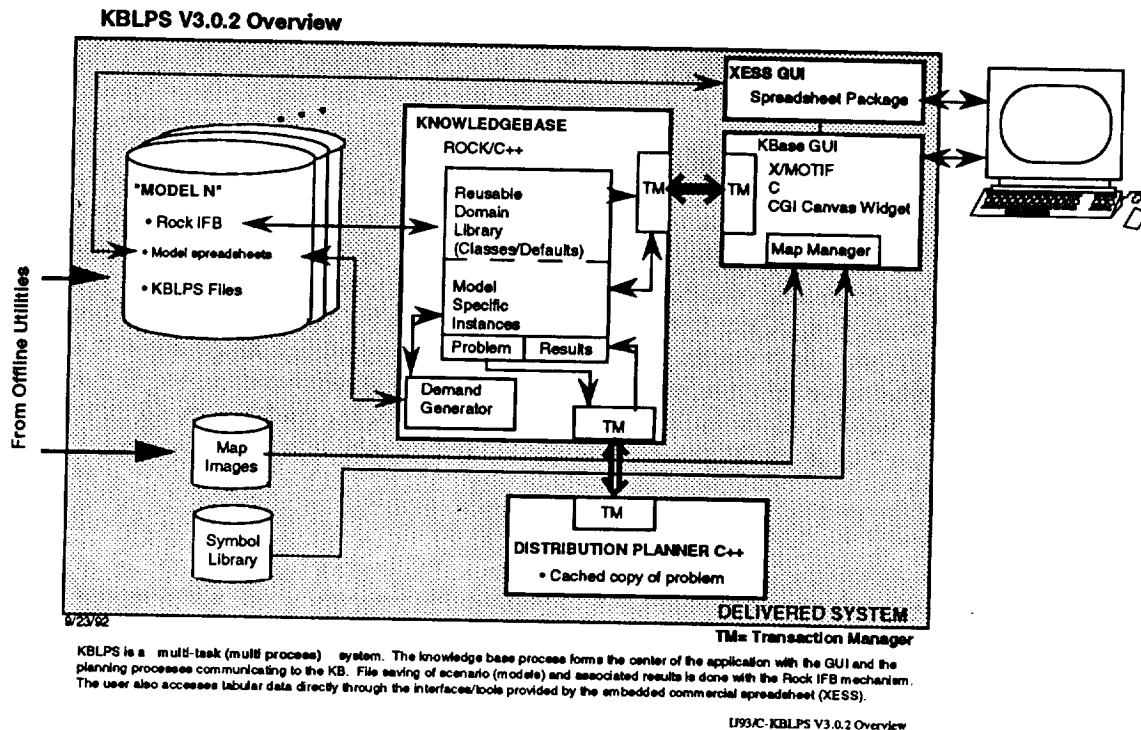
The complexity and dynamics of commodity distribution planning - both in commercial and military domains - require decision support tools that provide much more than data base access and spreadsheet solutions. This paper presents an overview of a Decision Support System (DSS) that incorporates Knowledge-Based (KB) and commercial off the shelf (COTS) technology components.

The Knowledge-Based Logistics Planning Shell (KBLPS) is a state-of-the-art DSS with a rich map-oriented interactive graphics user interface and powerful planning algorithms that has been designed and implemented to support skilled Army logisticians to much more rapidly prepare and evaluate logistics plans to support corps-level battle scenarios than currently possible. These plans may be developed as contingencies against future possible scenarios or in direct support of troops on the ground. In either case, the ability to build, evaluate, and improve plans in a fraction of the time now possible has been a major objective which is in the process of being met.

KBLPS is an appropriate blend of Artificial Intelligence, Knowledge-Based, conventional, and commercial off the shelf technologies that, taken together, provide a logistician with a powerful tool to help define, analyze, and evaluate very complex planning problems quickly. The Logistician can configure the problem/scenario with a through-the-screen object-oriented approach, as well as give guidance to the algorithm by (optionally) setting a number of parameters; the DP algorithm constructs distribution plans, involving significant computational complexity on behalf of the logistician. As a consequence, the logistician can spend more time analyzing and assessing plans than in generating them.

At the heart of KBLPS is the commodity Distribution Planner (DP) algorithm (see Figure 1), which supports Ammunition Distribution and bulk Petroleum Distribution planning in a single framework. A major challenge - one that has been met - was to design a single algorithm flexible enough for use in these two domain-specific areas. The DP algorithm reasons and calculates at appropriate levels of data aggregation to demonstrate the logistics supportability of a battle scenario without overwhelming the user with too much detail.

Figure 1:



KBLPS represents a substantial advance in graphical interactive planning tools, with the inclusion of intelligent planning algorithms that provide a powerful adjunct to the planning skills of distribution planners. KBLPS has been designed and implemented as a Decision Support System (DSS), based on the recognition and appreciation of how complex and challenging the logistics planning problem is, and how many subtle and interacting factors and considerations must simultaneously be brought to bear in solving these kinds of problems. The Logistician and the DSS are a team that works together, to achieve objectives that neither alone could accomplish as well. We believe that it is unlikely anytime in the foreseeable future that it will be feasible to replicate all the thought processes and decision making abilities of a skilled logistics planner; it is unlikely that would be desirable in any case.

Major benefits of using a DSS like KBLPS include:

- Faster Generation of Distribution Plans - Demand/Forecast Generation and DP algorithm generation of large-scale plans for huge 5-corps 3-5 day scenarios can be completed in 1-2 hours (computer execution in 5-10 minutes) compared to 24-48 hours turn around with currently manual procedures.

- Far greater data precision; KBLPS supports ammunition and petroleum flow to combat unit company and battalion level with 1-hour time interval granularity. Manual operations typically make estimates at division level using 24-hour time intervals.
- Far greater precision in taking into account time-varying availability of inter-acting resources. This degree of complexity simply cannot be addressed manually; there are too many interactions.

Rapid decision-making turn-around supports Army doctrine to make accurate decisions within the decision making cycle of the adversary. Being able to do this much more quickly and accurately with DSS aids - and enabling the consideration of many alternative "what-if" contingencies not now possible - meet major doctrinal goals.

PROBLEM DEFINITION

Figure 2 depicts the nature of the commodity distribution problem. A network of storage and distribution installations and transportation/material flow links are defined by the user in order to satisfy user commodity demands. In the particular application at hand, these are combat and combat service (CS, mainly artillery and engineering units) who are significant consumers of petroleum and ammunition. They could as easily be civilian consumers shopping at commercial retail outlets, which in turn are supported and fed by up-stream district and regional warehouses. Note that Figure 2 is merely representational; it is much simpler and easier to interpret and understand than real distribution networks with which skilled logisticians typically deal.

Figure 2:

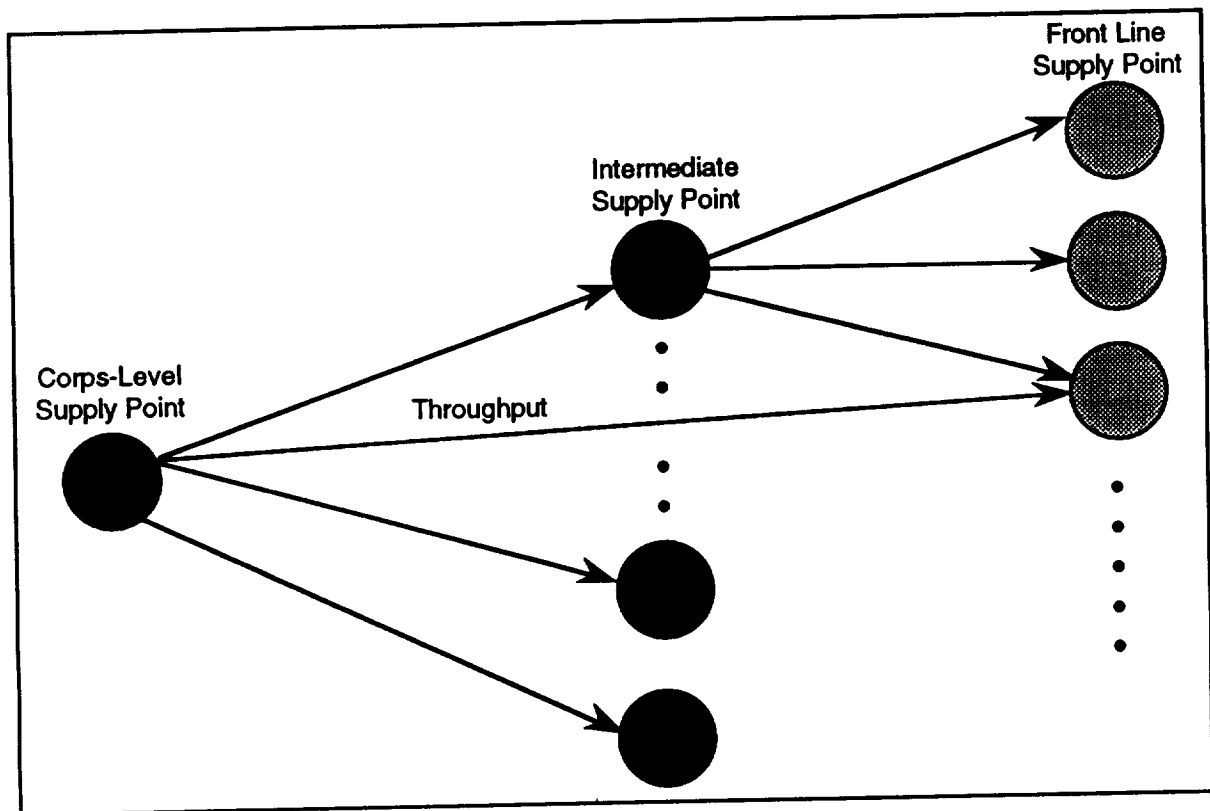
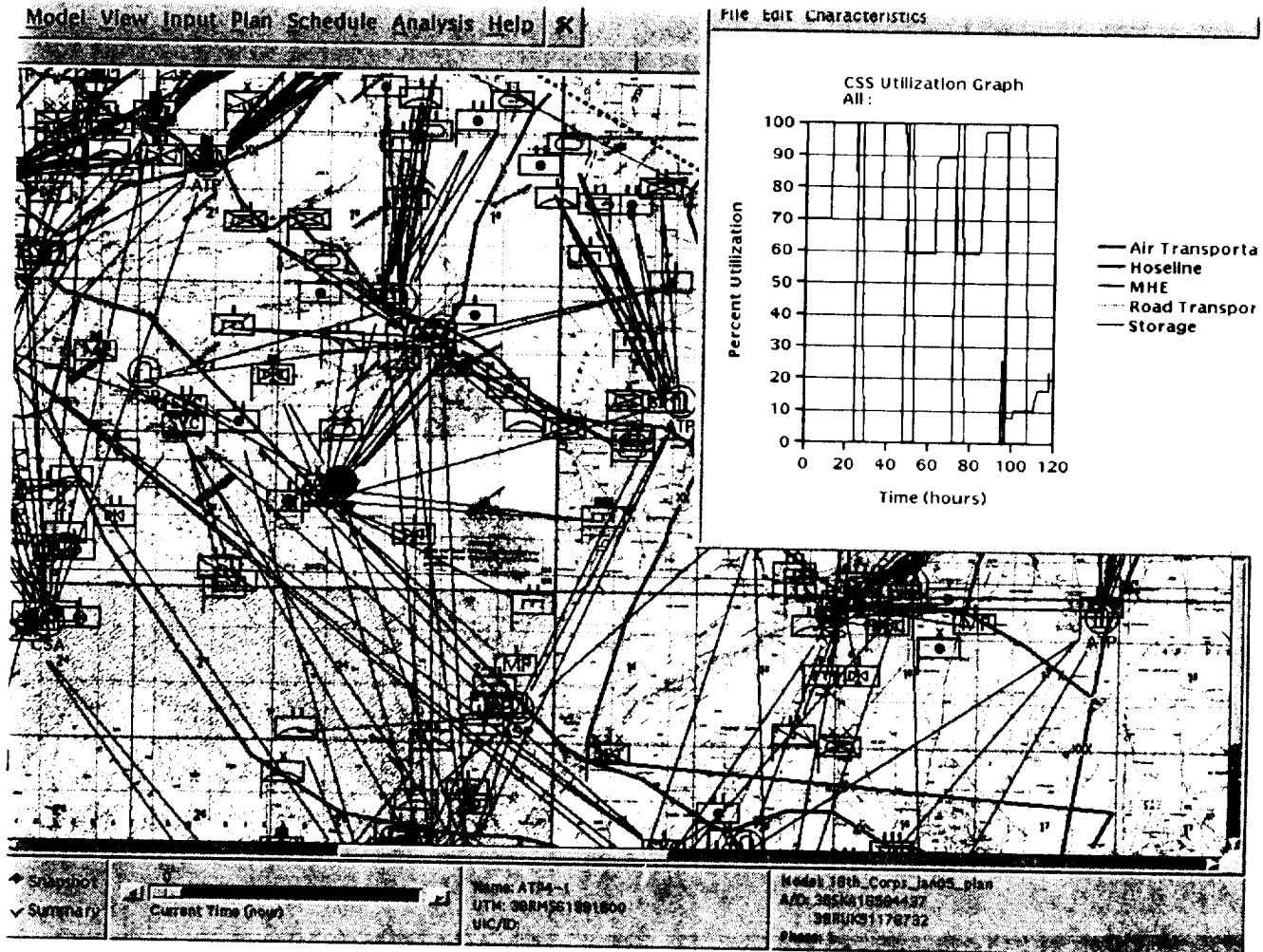


Figure 3 shows a typical screen view of a sector of the battlefield with Combat and CS units and their hierarchy links displayed. The black and white reproduction of the multi-color screen does not begin to do justice to the actual display, but is presented here as indicative of the richness and complexity of the display environment. The user has many controls to pick single 'layers' of displays, so that most screens viewed and used by the Logistician are far less complex and 'cluttered' than Figure 3.

Figure 3:



The Logistician's task is to configure the network, identify how much of each key resource (trucks, roads, material handling inventory), will be available (as a function of time), and assess whether that mix of resources will enable his customers (in this case combat and CS units) to do their job. In current practice with *situation maps* ("sit-maps") pinned to walls and transparent acetate overlays depicting various aspects of the problem (combat unit laydown, CS unit deployment, transport networks, etc.), logisticians perform this task with many approximations and heavy reliance on rules of thumb that may be inadequate or very difficult to adjust in new and unique circumstances.

KBLPS replicates the sit-map/acetate overlay paradigm electronically. The logistician can use through-the-screen commands to control which aspect of the planning problem to examine and manipulate. Because the displays are icon/object-oriented, the logistician can "click on and drag"

combat, CS, and Combat Service Support (CSS) units and installations to new locations on the underlying map in order to enhance the logistics network laydown. The KBLPS X-windows/Motif-based Graphics User Interface (GUI) provides the user with a complete and flexible environment which enables the Logistician to create and delete combat, CS, and CSS units at any echelon (level of command hierarchy), change support/supported relations, and change the capability and combat readiness of units to reflect actual or contingency conditions. A commercial spread sheet/graphics package was integrated into KBLPS to provide a complete and convenient way to manipulate the demand forecast model (unit structures, individual and roll-up to unit consumption rates as function of combat conditions, etc.), and to present summaries of generated plans from multiple perspectives.

KEY TECHNOLOGIES

Key elements of KBLPS are built on Knowledge Base/Artificial Intelligence foundations and state-of-the-art Graphics User Interface technologies.

(i) **The Knowledge Base (KB)** has been constructed using TM CGI's ROCK knowledge representation technology. ROCK provides a C/C++ implementation of KB functionality that enables application developers to build complex knowledge bases which capture complex and often *ad hoc* relationships. In this particular application there is a complex hierarchical command structure of an Army corps consisting of dozens of units (tens of thousands of troops), dozens of Combat Service (CS) units, and dozens of Combat Service Support (CSS) units and material storage/distribution installations. It was mandatory to use a KB approach in order to capture and effectively operate on such a complex and fluid domain, which includes extensive command hierarchy and support/supported relationships among different elements of the organization.

In addition, a KB approach provides significant flexibility in re-configuring a baseline (doctrinally defined) corps structure in light of particular combat situations that require some degree of command, support-supported, configuration and other modifications to the corps baseline. These same kinds of considerations apply as well in commercial commodity distribution networks.

(ii) **The Distribution and Transportation Planning Algorithms** apply a Constrained Heuristic Search (CHS) technique which analyzes key constrained resources (roads, pipelines, trucks, inventory, storage depots, material handling equipment) to build a *feasible* (i.e., no constraints violated) plan of forward material movement with rapid execution times on conventional engineering workstations.

CHS is a useful and powerful alternative to "conventional" approaches to solving complex planning, scheduling, and logistics problems. Techniques including Linear Programming (LP) and various forms of math programming, for example, have been used for many years in attempts to achieve optimal solutions, using formulated 'objective functions'. CHS provides a means to provide excellent (but sub-optimal) solutions that cannot be solved *quickly* (or at all) with LP, math programming, and other techniques.

CHS provides the opportunity to solve problems that cannot be formulated adequately in LP terms (given the significant degree of inherent non-linearity in the problems) or that would require an inordinate amount of time to compute a solution; by the time the solution is available, it may no longer apply when circumstances have changed in the meantime. The CHS formulation also makes it easier for the user to change the problem definition without having to rely on software engineers to re-code the problem.

This CHS/modeling approach is also fundamentally different from discrete event simulation (DES), another technique often used in planning and scheduling. The CHS model's view covers

the problem's entire time and geographical domain. Consequently, it can consider all up- and down-stream (in time and space) interactions while making its decisions. This is in contrast to DES, which is fundamentally a 'look downstream one step at a time' approach (although there are embellishments that can sometimes mitigate some of the shortfalls associated with DES), often leading to less than desirable results, and worse, usually does not provide the means to make changes to the problem statement/set-up in order to improve plan results.

(iii) **The Graphics User Interface (GUI)** has been implemented with X-windows/Motif layered on top of Government-furnished (electronic media) high quality maps. The GUI is an intuitively appealing and readily-learned means (as evidenced by how quickly skilled Army Logisticians have mastered the system) to set-up complex battle and logistics problems and evaluate and alter DP-generated plans. The GUI provides a convenient (control mouse) point-and-shoot approach to modify force and logistics deployment configurations, create or delete units at any echelon level in seconds, modify unit capability profiles in seconds, etc. The ability to see the entire problem at a glance and to be able to home in on particular elements of the problem quickly and easily is a key component to the User/DSS shared problem solving paradigm that KBLPS represents.

PLANNING OBJECTIVES & APPROACH

The purpose of the Distribution Planner (DP) algorithm is to help the Logistician prepare distribution plans that simultaneously meet these goals:

- Meet required delivery dates and quantities on time, or somewhat early ("Just In Time" delivery)
- Fill highest priority needs first
- Balance service across units (the customers)
- Balance service of high, medium, and low priority needs
- Maintain stockage objectives at distribution installations
- Make efficient equipment, transport, and supply route assignments

An important reality is the need to plan against multiple goals or objectives (some of which are conflicting) in a data-intensive, non-linear, time-varying context. For example, the goal to maintain certain time-varying stockage objectives may conflict with the goal to reserve 20% of main supply route (MSR) capacity for contingencies. Logistics resources are usually highly constrained; i.e., there is usually much more demand for ammunition and petroleum at the front lines than can be supplied by the available or projected distribution assets. The objective is to provide:

- Feasible robust (though not necessarily optimal - it is not at all clear how to define 'optimal' in this context to everyone's or anyone's satisfaction) plans quickly,
- Information and insight to the Logistician that will help guide the distribution planning process, satisfying these multiple goals in as even-handed and balanced a way as possible,

- Insight to the Logistician and the commander where the plan has the most risk and how such risks might be mitigated.

The CHS-based algorithm cycles through the problem domain, step-wise building up a distribution plan across a user-defined time and geographic horizon. The CHS algorithm looks at the whole problem all of the time (this distinguishes it from dispatch-based planning and event-driven simulation approaches, for example), successively turning its focus to those resources at particular time intervals most in need of resolution before dealing with less-constrained and less urgent problem elements.

The DP algorithm constructs a plan by iteratively selecting an order (the 'next best' order not yet in the plan) and then placing that order into the plan. The DP reasons and plans "opportunistically", i.e., neither the sequence in which orders are planned nor the way in which orders will be planned are determined in advance. Since there are typically many orders which could be planned next, and numerous ways to plan them, the DP uses heuristics to guide the search to make these decisions quickly, efficiently, and accurately. Hence, the DP algorithm is fundamentally data driven, responding 'opportunistically' to different problem scenarios, 'floating bottlenecks', etc., thereby avoiding the need for the user to intervene (other than with input data modifications as appropriate) when problem circumstances change.

DETAILED DP ALGORITHM DISCUSSION

The DP algorithm can plan the distribution of commodities through a distribution network with sufficient flexibility to apply to both petroleum and ammunition distribution. In general, this problem is usually over-constrained - there are not enough assets (material handling equipment (MHE), trucks, main supply route (MSR) capacity, inventory) to satisfy all demand. Moreover, in different scenarios different resources will be most scarce. Thus, the algorithm must maximize efficient usage of all resources, and in particular the scarce resources, in order to maximize over-all order satisfaction. This must be done in light of commander intent/mission, balanced servicing of all combat units so they can adequately support the mission and each other, and available or anticipated assets.

The principal goal of the DP algorithm is to maximize the satisfaction of Combat, CS, and CSS unit demands for petroleum or ammunition. However, priorities must be respected so that higher priority orders are satisfied before lower priority orders. But the preference for higher priority orders must be tempered by the objective that lower priority orders cannot be completely ignored and starved; hence, there must be a balance between these conflicting goals. The DP algorithm provides a mechanism which allows a Logistician-specified percentage of higher priority orders to be satisfied first, then a percentage of lower priority orders, and then the remainder of the higher priority orders. In this way, priority is respected but lower priority units and orders are not starved. This balancing mechanism gives control to the Logistician, and is consistent with the KBLPS philosophy of being a Decision Support System.

An important secondary objective is to meet day-to-day inventory stockage targets, which are Logistician-specified in terms of number of days of supply in the context of the battle scenario being supported. When in conflict with the primary objective (i.e., service at the front line/'point of sale'), this objective is relaxed.

The DP must assign MHE, MSR, truck and available inventories to supply the user units consumption/demands in the best possible way over time. In general, this is an "over-constrained" problem - the net demand is generally substantially greater than inventory and/or material handling and carrying capacity can handle, at least in some time-frames. Hence, some of the demand will remain unsatisfied. It is the Logistician's task, with the aid of the DP, to find the means to move

as much material forward as possible and to set aside that subset of demand that can best be 'held back' and still meet the commander's intent. Procedurally, the Logistician uses the DP iteratively, successively refining the scenario definition and data elements, to move toward better and better scenarios.

Initial Use & Current Status

KBLPS has been undergoing early user evaluation by skilled logisticians of the XVIIIth Airborne Corps, Ft. Bragg, NC, over the past year. KBLPS has been used and stress-tested in the context of a series of large scale realistic logistics and force projection exercises at a number of sites and in diverse scenarios (including Prairie Warrior exercise at Ft. Leavenworth, KS, in May 1993, and Force Projection Logistics Exercise at Ft. Lee, VA, in July 1993). The speed with which skilled logisticians have learned to use KBLPS has been especially noteworthy. It reflects the degree to which the users know their business in depth and how well the KBLPS implementation matches the way they think about and solve their planning problems.

Further extensions to KBLPS are under development. These include: incorporation of other classes of supply - food, medical supplies, major end items, spare parts; enhanced plan analysis tools; integration of the more-detailed *Transportation Scheduler* with the 'rough cut' Distribution Planner, etc. Future plans include moving toward an on-line reactive re-planning capability to support 'current operations' (vs. look-ahead contingency planning) and cutting-edge visualization techniques, including such approaches as 2 1/2-D perspective displays and plan animation.

Potential Application to Commercial Domain

We believe that KBLPS can be readily applied to strategic planning in commercial environments. Rapid and effective evaluation of the logistics impact of a corporation's rapid growth or moving into new markets or introducing new product lines is a clear strategic need and opportunity. Other contexts would include mergers and acquisitions, providing third-party logistics/distribution services to other corporations, and contingency/disaster response planning. The basic JIT paradigm referred to earlier, the underlying flexibility of the "Activity Model" (process plan) implemented within the algorithm, and the richness and flexibility of the User Interface, are key elements to a straightforward transition into the commercial arena.

Some important extensions to KBLPS that we anticipate could be useful or required include changing the demand forecast model from a military planning factors structure to more conventional forecast models often provided by MRP systems, for example; changing time granularity from hours (currently) to days, weeks, or months; adding new modes of transport (sea, waterway) - straightforward, given the Activity Model paradigm; and extending the stockage/resupply model to accommodate stock carrying costs and re-order points which differ from the current inventory model. Map display quality and granularity could probably be relaxed substantially from the current implementation.

SUMMARY

The Knowledge Based Logistics Planning Shell is a cutting-edge Decision Support Aid which combines the best of conventional (X-windows, Motif, commercial spreadsheet/graphics package, etc.) and Artificial Intelligence (Knowledge Base and constraint-directed planning algorithms) technologies. There is substantial potential for extensions, including platforming new applications modules on top of the extensive knowledge base now in place.

We have designed and implemented a flexible algorithmic approach that quickly (3 - 5 minutes execution speed on conventional work-stations) provides feasible plans accurately for large

problems (2,000+ orders over 120 time intervals, dozens of customers and distribution points, hundreds of constrained resources).

This Decision Support System is currently in active use by skilled logisticians, supporting their on-going efforts to plan for regional contingencies in many parts of the world. The users', and their senior management's, current enthusiasm and support have been key elements in guiding on-going development.

We gratefully acknowledge the support and sponsorship of
Mr. Richard Camden, U.S. Army Research Laboratory,
and the U.S. Army Strategic Logistics Agency.

BIBLIOGRAPHY

1. M. S. Fox, Observations on the Role of Constraints in Problem Solving, In Proceedings of the Annual Conference of the Canadian Society for Computational Studies of Intelligence, May 1986
2. M. S. Fox, N. Sadeh, C. Baykan. Constrained Heuristic Search. In Proceedings IJCAI-89, 1989
3. M. S. Fox and S. F. Smith, ISIS: A Knowledge-Based System for Factory Scheduling, In Expert Systems 1,1(1984)
4. P. S. Ow and S. F. Smith, Viewing Scheduling as an Opportunistic Problem-Solving Process, In Annals of Operations Research, 12 (1988)
5. N. Sadeh and M. S. Fox. Variable and Value Ordering Heuristics for Activity-Based Job-Shop Scheduling. In Proceedings of the Fourth International Conference on Expert Systems in Production and Operations Management, 1990
6. N. Sadeh. Look-Ahead Techniques for Micro- Opportunistic Job Shop Scheduling, Ph. D. Thesis, CMU, 1991
7. M. Zweben, M. Dale, R. Gargon. Anytime Rescheduling. In Proceedings of the DARPA Workshop on Innovative Approaches to Planning and Scheduling, 1990
8. V. Saks, A. Kepner, I. Johnson. Knowledge Based Distribution Planning, American Defense Preparedness Association Conference, March 30, 1992, Williamsburg, VA.
9. V. Saks, I. Johnson. Constraint Directed Distribution Planning: Lessons Learned, American Defense Preparedness Association Conference, March 7, 1993, Williamsburg, VA.

A HYPERTEXT SYSTEM THAT LEARNS FROM USER FEEDBACK

Dr. Nathalie Mathé
NASA Ames Research Center¹
Mail Stop 269-2
Moffett Field, CA 94035
mathe@ptolemy.arc.nasa.gov

2491
p. 7

ABSTRACT

Retrieving specific information from large amounts of documentation is not an easy task. It could be facilitated if information relevant in the current problem solving context could be automatically supplied to the user. As a first step towards this goal, we have developed an intelligent hypertext system called CID (Computer Integrated Documentation). Besides providing an hypertext interface for browsing large documents, the CID system automatically acquires and reuses the context in which previous searches were appropriate. This mechanism utilizes on-line user information requirements and relevance feedback either to reinforce current indexing in case of success or to generate new knowledge in case of failure. Thus, the user continually augments and refines the intelligence of the retrieval system. This allows the CID system to provide helpful responses, based on previous usage of the documentation, and to improve its performance over time. We successfully tested the CID system with users of the Space Station Freedom requirement documents. We are currently extending CID to other application domains (Space Shuttle operations documents, airplane maintenance manuals, and on-line training). We are also exploring potential commercialization of this technique.

INTRODUCTION

There is a need in NASA, and other governmental and commercial companies for rapid and effective access to information in large documentation systems. Accessing and using information stored in such documentation systems has become more and more difficult due to the increasing volume of documentation, the disparity of information sources, and the frequent rate of change of such documentation during its lifetime. One of the main problems in accessing information in large documentation systems is due to the lack of support for contextual information retrieval. The relevance of information not only depends on its contents, but also on the particular user and his/her current problem solving context.

Hypertext systems provide good support for browsing large amounts of documentation and for building associative links between various parts of the documentation. While hypertext systems increase accessibility of information, they do not provide any built-in selectivity mechanism. This increased accessibility may magnify an already severe problem of selection [1]. For these reasons, knowledge-based systems technology can be very helpful in alleviating the selection problem and the cognitive overhead of the user. Our approach on the CID project has been to develop an intelligent context-sensitive indexing and retrieval mechanism which interacts with and learns from the user [2]. This mechanism lets users filter information by context of relevance, and build personalized views of the documents over time. It also gives them the capability to share knowledge with other users.

RETRIEVING INFORMATION IN LARGE DOCUMENTS

We have analyzed uses of the Space Station Freedom (SSF) Program Requirement Documents. When looking for specific information, users usually employ both the table of contents (hierarchical browsing) and the index of the available documents to guide their search. Even then, however, the search is not very focused, and it is common for users to examine several places in the documentation before finding the information needed. Moreover, the needed information might be distributed in several locations, making it harder to exhaustively find all the relevant pieces. In other cases, the information might not be in the document at all, but the user has to be certain of it before requesting a change to the document.

¹ Dr. Mathé is currently working at NASA ARC under contract 21-1614-2360 to San Jose State University.

While in a small document the user could use a simple sequential trial-and-error decision process (go to each section mentioned in the index and read its text), this is not feasible in large documents. We observed that users prefer to avoid reading pages of text, and apply an iterative filtering strategy: select a rough set of candidate locations in the document, use high level information to decide which ones might be relevant, repeat this process until only a few locations have been identified, then access and search the text at each location. If the information is not found, or only partial information is found, then backtrack in the search by broadening the set of candidate locations. In such a strategy, reading text is used in the last resort.

To avoid repeating this lengthy process each time, users develop cognitive representations of the organization and content of the documentation. They build semantic descriptions of the locations already visited in the text, and contextual information around the descriptors already used to access particular pieces of information. These cognitive maps are thus context- and user-dependent. This memory however does not last very long unless the same information is retrieved often under the same context.

THE COMPUTER INTEGRATED DOCUMENTATION SYSTEM

Retrieving specific information from large amounts of documentation could be facilitated if information relevant in the current problem solving context could be automatically supplied to the user, in understandable terms and in a flexible manner. As a first step towards this goal, we have developed an intelligent hypertext system called CID (Computer Integrated Documentation) [3]. The CID system enables integration of various technical documents in a hypertext framework and includes an intelligent context-sensitive indexing and retrieval mechanism. The CID system has been used to help Space Station Level I personnel manage the Program Requirement Document for SSF. A typical screen of CID windows is presented in Figure 1. It contains the CID control panel that allows the user to control the entire library, and an example document. Both text and graphics capabilities are available within CID.

Structured Technical Documents

In existing technical documents, such as those common at NASA, information is structured hierarchically. Designing a complex system like the Space Station Freedom is an iterative process. Its documentation system is designed to handle a huge amount of information. It is organized around the Program Requirement Document (PRD), which establishes the highest level requirements associated with the Space Station Program. Generally, lower level documents expand upon the topics expressed in the PRD. For instance, Figure 1 shows the CID version of the PDRD (Program Definitions and Requirements Document, one level below the PRD) Section 3 (Systems Requirements), which contains about 900 pages, and uses 1.4 Mb of memory (without figures).

Each document includes several major nodes (e.g., change and revision notices, table of contents, a body of text segmented into sections and subsections, tables, figures, various appendices.) There are references between sections, or links in the hypertext "language", which are linear (section to next section) and non linear (reference to a section other than the next one). There are references to other major nodes within a document and to other documents.

In CID, we call *descriptor*² any word, phrase, or piece of graphics which provides a meaningful "starting point" for a search in the documentation. A *referent* is the address of any part of text or graphics. In the current implementation, a descriptor is typically a word in the index, and a referent a document section. Initial hypertext links between descriptors and referents are automatically built by CID, and may be modified by the user. CID also lets users index referents with concepts that are not necessarily words or term-phrases included in the text.

² The concept of descriptor is very important in information retrieval [4]. Building such descriptors requires expertise in the domain of investigation. We are using a technique developed by Mark Zimmerman [5] that allows full-text extraction of words associated with their frequency in the text.

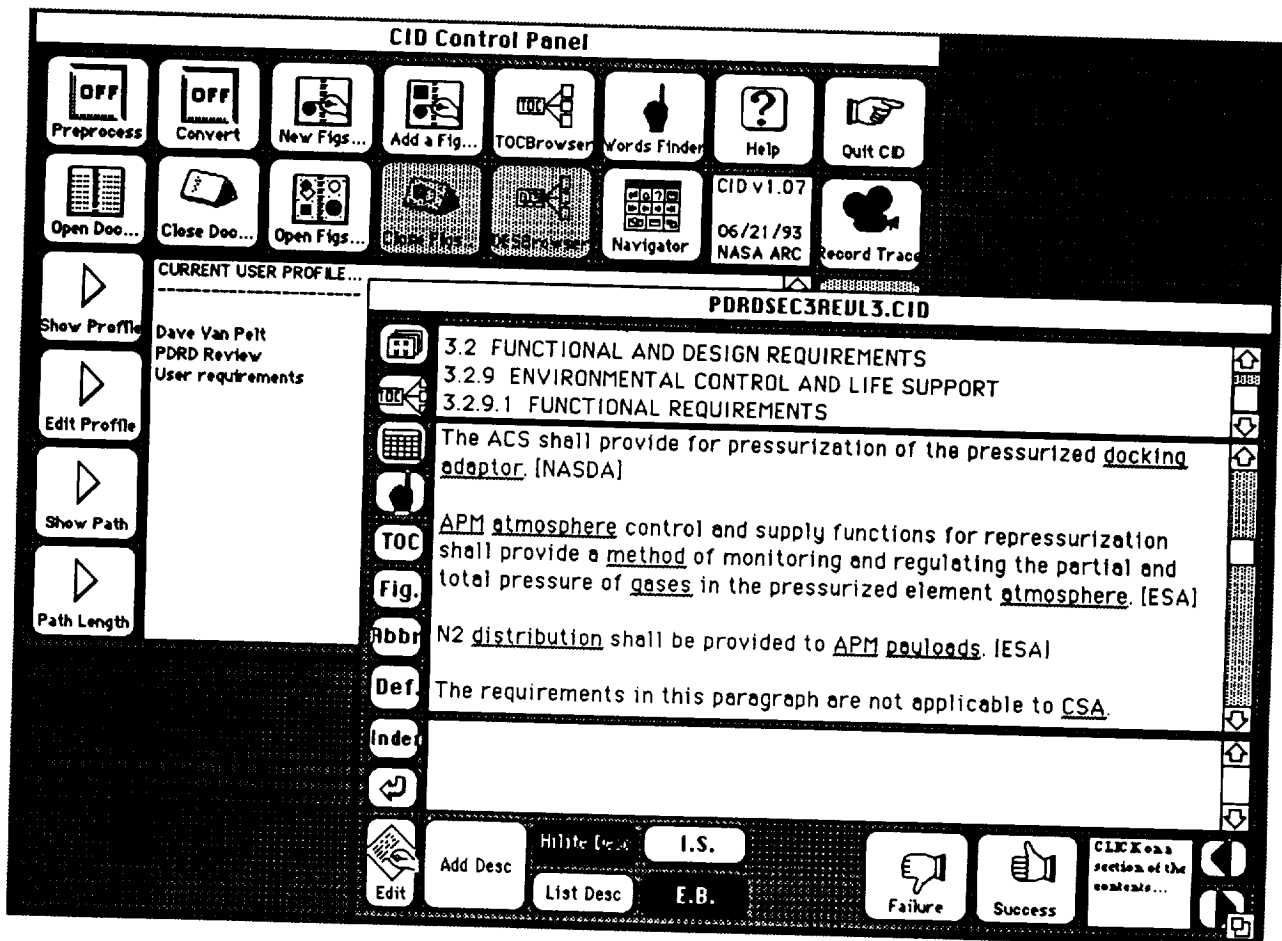


Figure 1 Part of a typical CID screen. The upper-left window is the CID control panel. The lower-right window is a CID document. The upper field includes the hierarchy of the displayed referent, the middle field is the actual text and the bottom field includes the next referents in the hierarchy. Clicking on the Success or Failure button automatically reinforces the displayed referent under the current user profile.

Context-Sensitive Retrieval

Contextual information characterizes the user, the types of task he/she is usually performing, and the type of information he/she is looking for. For instance, let us assume that one wants to retrieve some very specific information on the disposal of waste water in the Space Station. The first thing one may try is to browse the documentation with the descriptor "waste water." If the retrieval context can be specified, e.g., "You are a future scientific user of Space Station, and you want to know how experimental waste water will be handled," then a more efficient search can be accomplished. The set of relevant referents will not be the same under another context, e.g., "You are an engineer and you want to know how waste water will be recycled in the Environmental Control and Life Support system." Currently, in CID, the retrieval context is set by default on user log-on to one of the user's profiles, and can be modified by the user at any time.

Retrieving information in documentation is generally handled using keyword search. People find this very difficult in practice because keywords are used in a full-text search mode. Consequently, systems using keyword search come up with either hundreds of references or nothing [4]. To overcome this selection problem, CID uses the current retrieval context to filter down the number of referents for a descriptor that a user has to look at and thus narrows down the search. When the user selects a descriptor in CID, a list of ordered referents pops up. These referents have been found successfully in the *same or similar context* in past retrievals (the user can, however, access all the existing referents at any time). The order of referents is based on the past success and failure rates of each referent in this context. This is illustrated in Figure 2: the number of referents retrieved using the context filtering capability is narrowed down from 26 (total number of existing referents) to 2.

Users have the choice between filtering the list of retrieved referents based on previous reinforcements they have done themselves, or based on reinforcements done by someone with a profile similar to their own current profile. This second filtering option lets users with similar interests share their knowledge about the relevance of various pieces of information in a document. In the same way, a novice user can immediately visualize reinforcements done by a domain expert on the document, and shorten his/her training time.

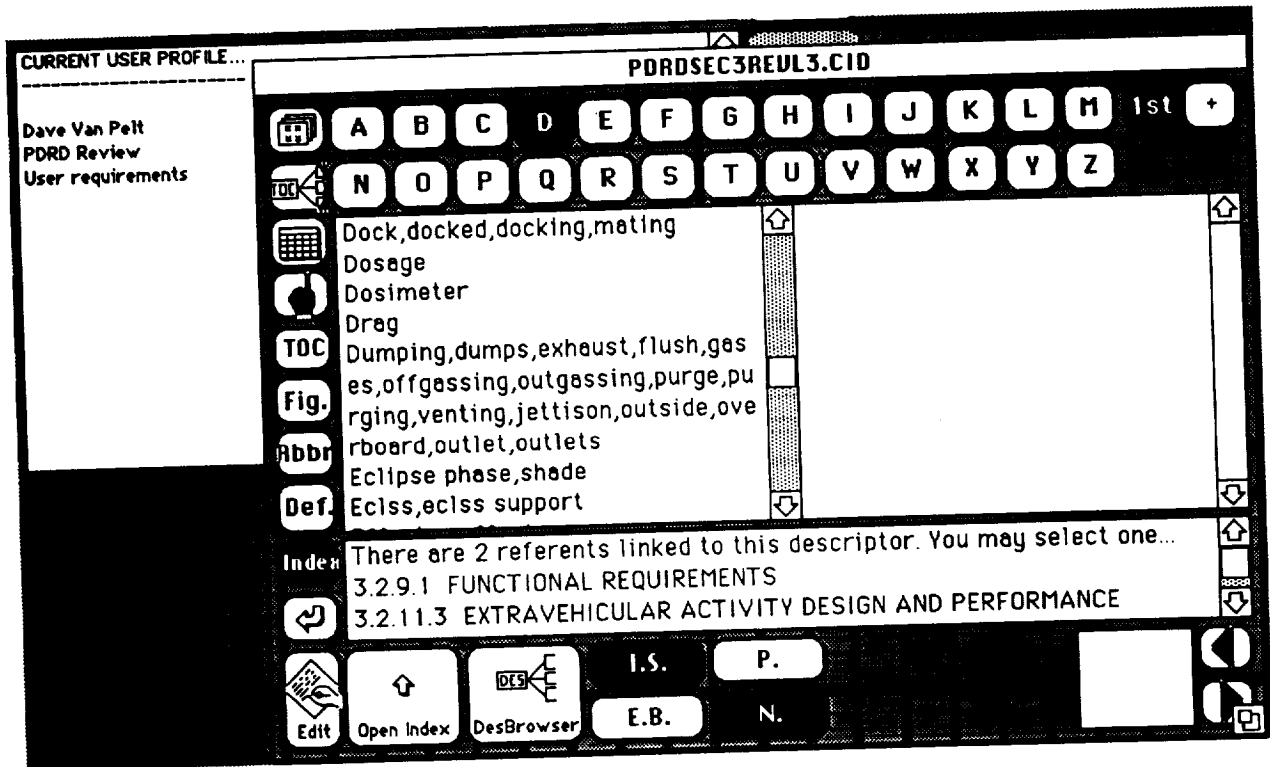


Figure 2 Use of an index in context. Here, the context was characterized by the current user profile. The user clicked on the descriptor "Dock". CID suggested two referents previously reinforced under this context.

Context-Sensitive Reinforcement

In order to narrow down the search using the current retrieval context, CID automatically acquires contextual knowledge from previous searches. This contextual knowledge is associated to the indexing links between descriptors and referents. CID modifies existing relations between descriptors and referents by using on-line user feedback to either reinforce or correct the system's knowledge in case of success or failure. This feature allows the system to tailor itself to the user.

After accessing a referent from a particular descriptor, the user can select either "Success" or "Failure" (cf. Fig. 1) to specify the relevance of this referent to his/her search. Using the current retrieval context, CID automatically updates the contextual knowledge attached to this link, updating contextual conditions and their respective weights (cf. Fig. 3). When a failure occurs, the system attempts to obtain from the user the reason for this failure. This learning mechanism enables CID to refine the indices over time to reflect their context of use.

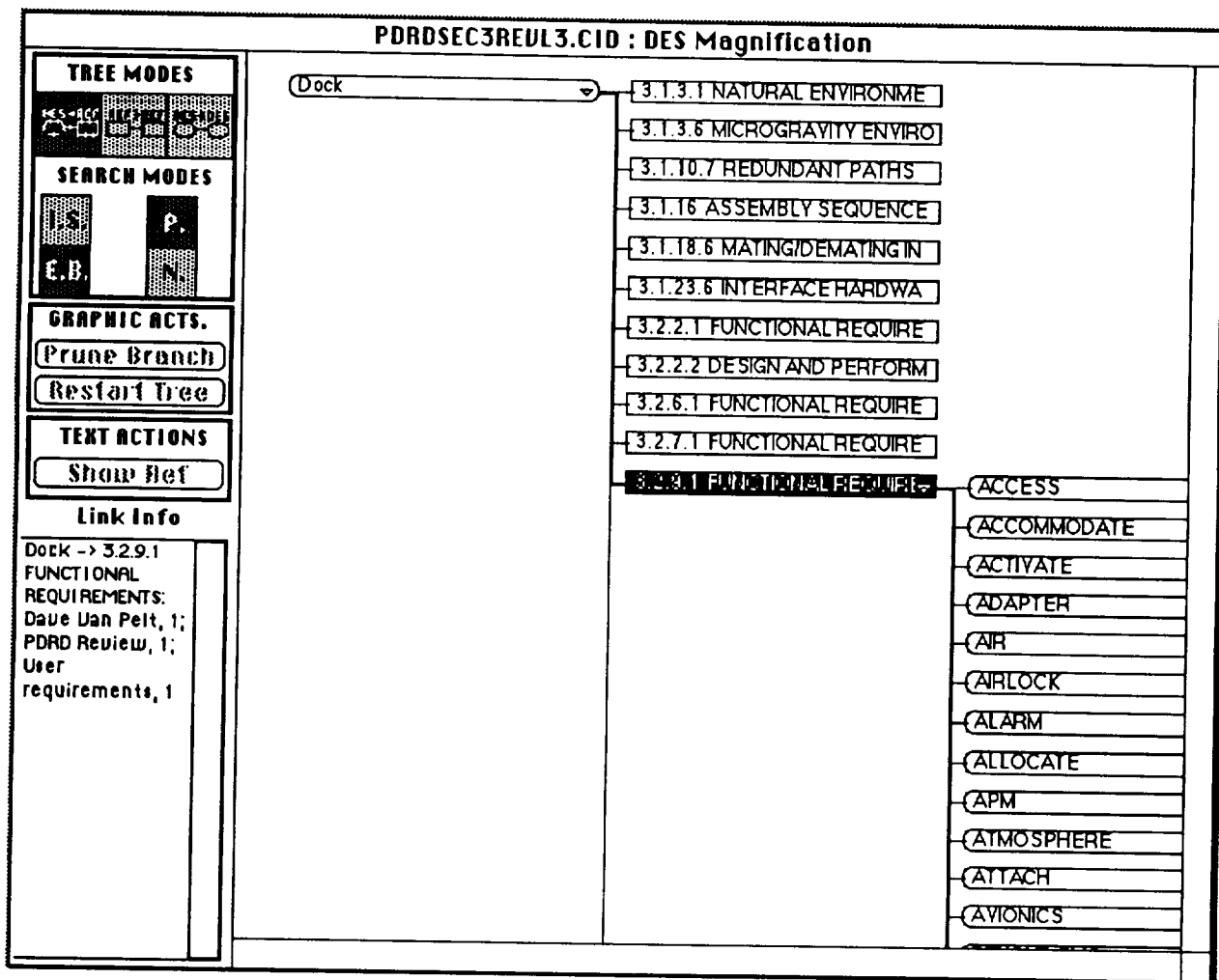


Figure 3 The graphical descriptor browser lets users visualize reinforcements (bold lines) and browse indexing links. Descriptors are displayed as rounded nodes, and referents as rectangular nodes. Contextual conditions associated to the selected link are displayed in the lower left corner. Clicking on a referent node displays the text of this section in the CID document.

CID Graphical Browsers

We have found that explicit maps of the documentation are very useful for quickly accessing information. These maps can be local ("where to go next?"), or global ("where am I?"). They can also present either the hierarchical structure of the documentation (local or global tables of contents), or the conceptual relationships between referents via descriptors (local or global conceptual indexes). In CID, these maps are embedded directly in the hypertext or accessible through graphical browsers.

Each section in a document locally displays information about its parent and children sections, and users can directly visualize in the text the words under which this section is indexed (cf. Fig. 1). These underlined words are active, and let users jump to other sections indexed under the same words.

User also have access to global maps displayed in graphical browsers. The graphical descriptor browser shown in Figure 3 lets users visualize reinforcements (bold lines) and browse relationships between referents and descriptors. In addition to browsing indexing links, they can ask for all referents similar to a given referent (sharing some descriptor with it), and filter the resulting list by context. The graphical table of contents browser lets users visualize the hierarchical structure of a document (cf. Fig. 4). In both browsers, all nodes are expandable and collapsible, and any section can be displayed directly in the corresponding document by clicking on a referent node.

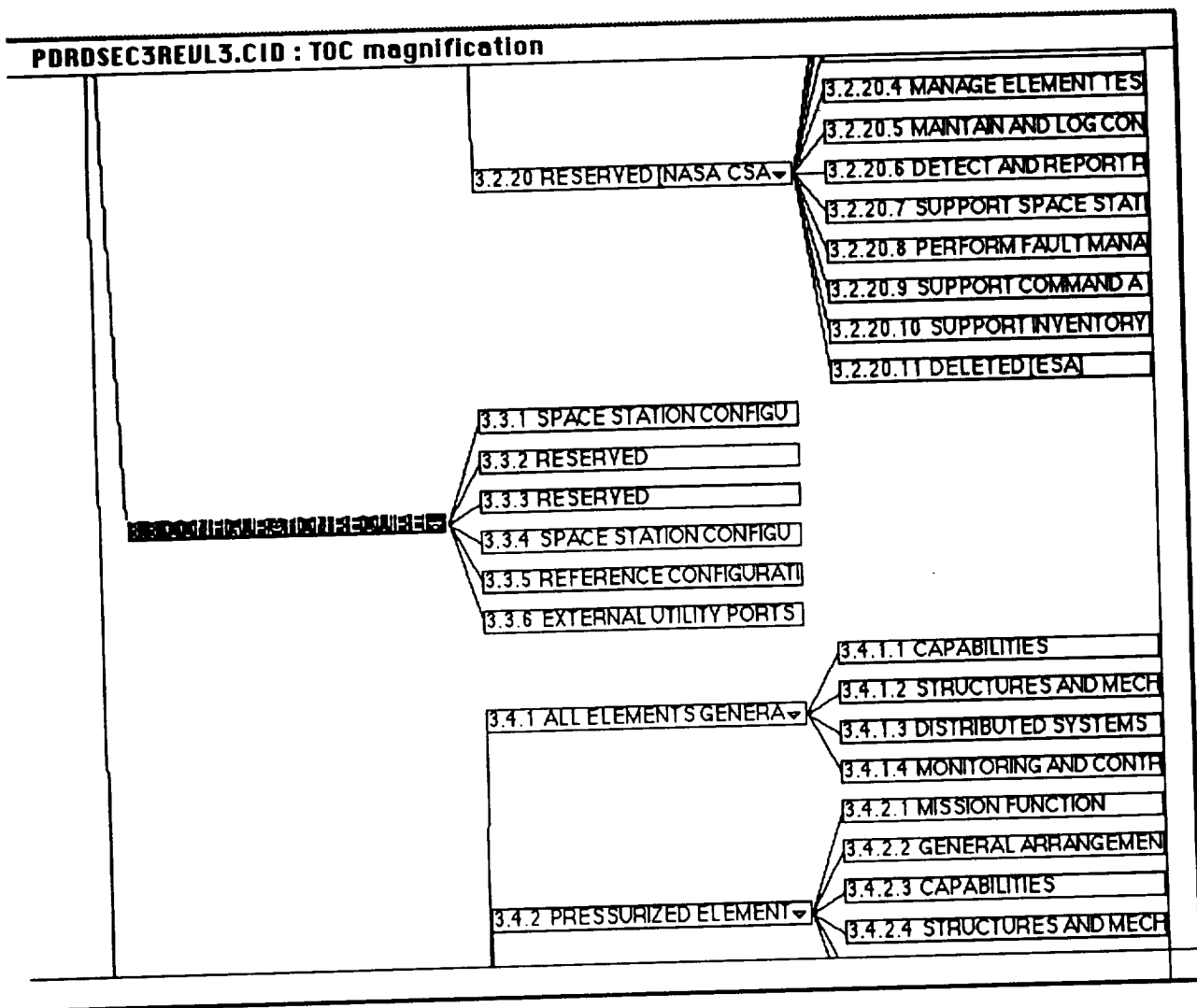


Figure 4 The graphical Table of Contents browser lets users visualize the structure of a document. Nodes can be expanded and collapsed. Clicking on a node displays the text of this section in the CID document.

USER TESTING RESULTS

In order to evaluate the effectiveness of and user satisfaction with CID learning capabilities, we conducted a set of user testing experiments [6]. Two domain experts from Space Station Level I personnel at NASA HQ used CID to find information in the Space Station Program Definition Requirements Document (PDRD). Their main task is to answer queries from future scientific users of the Space Station, regarding whether the current requirements fit the scientists' experiments needs. We chose 15 queries received by these experts from scientists, and not answered yet. Users were trained the previous week on CID for about half an hour. During each test session, we automatically recorded all their actions in CID, and asked them to verbalize what they were doing (think aloud protocol).

Overall, users found CID very user-friendly and were very eager to use the relevance feedback mechanism. In fact, they used it almost systematically each time they found some information they were looking for. With paper documents, they usually put yellow stickers and highlight the text with a yellow marker, in order to later on compose a report from the various pieces of information found in the document. With CID electronic documents, users only need to set up their user profile once (containing information to be used to personalize their feedback), then to click on the Success or Failure buttons whenever they want to give relevance feedback. Personalizing relevant information with their user profile also lets them share information with other users.

Apart from these positive results, users also requested more capabilities to make a better use of the CID system, including: better visualization of reinforcements directly in the document (highlighted sections, index terms, etc.); being able to reinforce any references, and not only those accessed from the index search; better filtering and access to previous reinforcements; being able to save reinforcements done for a query (which may involve several index searches or other types of searches) under one label; and means for transferring reinforcements done on a document to new revisions.

We are currently redesigning CID reinforcement mechanism in order to satisfy these new requirements. In this new version, users can reinforce any type of hypertext link, and have access to a broader type of queries, e.g., "show me all reinforcements I have done", "show me all reinforcements John has done for this query", "show me all reinforcements done yesterday", etc.

CONCLUSION

Besides providing an hypertext interface for browsing large documents, the ability of our system to automatically acquire and reuse the context in which previous searches were appropriate is unique. The design of contextual links to retrieve information is based not only on the way the documentation has been built, but also on user's information requirements and feedback when they are using the system. Thus, the user continually augments and refines the intelligence of the retrieval system. Context-sensitive information retrieval gives extended possibilities such as providing search expertise from other users, e.g., "what would John Smith do in this situation?"

We have shown that tailoring of hypertext documents during usage has been very well received by users. The main advantage is that it let them incorporate their knowledge and understanding of the content of a document over time, without disturbing the task they have to accomplish. This mechanism is extremely useful for large documents used by a large number of users, where the need for filtering information and building personalized views of the documents over time is important, as well as the possibility to share knowledge between users (access tailoring done by someone else, or a group of users).

Besides improving CID reinforcement mechanism as described in the previous section, we are extending CID to support operations at the Consolidated Control Center for Space Shuttle and Space Station in collaboration with the NASA Johnson Space Center. We are also studying the usefulness of CID for airplane maintenance electronic performance support with on-line documentation and training. Finally, we are exploring the potential commercialization of CID learning technique with computer software companies.

The current version of CID is implemented on a Macintosh in HyperCard and in C language. The CID software package is available for distribution to US universities and companies.

REFERENCES

1. Jones, W.P., "How do we distinguish the hyper from the hype in non-linear text ?," in *Proc. 1987 INTERACT'87*. . Holland: Elsevier Science, 1987.
2. Boy, G.A., "Acquiring and refining indices according to context," in *Proc. 1990 Fifth AAAI-Sponsored Knowledge Acquisition for Knowledge-Based Systems Workshop*. . Banff, Canada: November 1990.
3. Mathé, N. and G.A. Boy, "The Computer Integrated Documentation Project: A Merge of Hypermedia and AI Techniques," in *Proc. 1992 Space Operations Application and Research (SOAR) Workshop*. . NASA Johnson Space Center, Houston, Texas: August 3-6, 1992.
4. Salton, G., *Automatic Text Processing: the transformation analysis, and retrieval of information by computers*. Redding, Ma: Addison Wesley. 1989.
5. Zimmerman, M., *TEXAS version 0.5*, Technical Report. Silver Spring, MD: 1988.
6. Mathé, N., "Tailoring Hypertext Documents in Context: First User Testing Results," Poster presented at the 1993 Hypertext conference. Seattle, Ca: Nov. 14-16, 1993.

omit

COMPUTER-AIDED DESIGN AND ENGINEERING

COMMON MODELING SYSTEM FOR DIGITAL SIMULATION

Capt Rick Painter, USAF
USAF Wright Laboratory/Avionics Directorate
Wright Patterson AFB OH 45433

ABSTRACT

The Joint Modeling and Simulation System is a tri-service investigation into a common modeling framework for the development of digital models. The basis for the success of this framework is a X-window-based, open systems architecture, object-based/oriented methodology, standard interface approach to digital model construction, configuration, execution, and post processing.

For years Department of Defense (DoD) agencies have produced various weapon systems/technologies, and typically digital representations of those. These digital representations (models) have also been developed for other reasons such as studies and analysis, Cost Effectiveness Analysis (COEA) tradeoffs, etc.

Unfortunately, there have been no Modeling and Simulation (M&S) standards, guidelines, or efforts towards commonality in DoD M&S. The typical scenario is an organization hires a contractor to build hardware, and in doing so a digital model may be constructed. Until recently, this model was not even obtained by the organization. Even if it was procured, it was on a unique platform, in a unique language, with unique interfaces, and, with the result being UNIQUE maintenance required. Additionally, the constructors of the model expended MORE effort in writing the "infrastructure" of the model/simulation (e.g. user interface, database/database management system, data journalizing/archiving, graphical presentations, environment characteristics, other components in the simulation, etc.) than in producing the model of the desired system. Other side effects include: duplication of efforts; varying assumptions; lack of credibility/validation, decentralization both in policy AND execution, and various others. J-MASS provides the infrastructure, standards, toolset, and architecture to permit M&S developers and analysts to concentrate on the their area of interest.

J-MASS ARCHITECTURE and STANDARDS LAYERS

J-MASS has several architectural and standardization layers. This paper describes J-MASS in terms of the Tool Interconnect Backplane (IBP) layer, referred to as the Simulation Support Environment (SSE IBP) the Simulation Runtime Agent (SRA) IBP layer, and the Model Component/Object Standards layer.

MODEL COMPONENT/OBJECT STANDARDS

Each model component (or object) in J-MASS is structured compliant with our Software Structural Model (SSM). The SSM evolved from the Software Engineering Institute (SEI) work on the Object Connection Update (OCU) model. Both the C-17 and B-2 weapon systems trainers use a similar methodology for their object definition. The SSM, also described in a document, enforces software structure and interface standards for all levels of object decomposition. In this way, ANY objects in the system can be syntactically "connected" with any other objects in the system with guaranteed success. Semantically, the connection may have no realistic "meaning", but syntactically they can be connected ("Assembled", see discussion in Develop and Assemble Modes under Tool Interconnect Backplane). J-MASS objects are described in three layers: "Players", "Assemblies", and "Elements". Players are the "top" level objects responsible for synchronization with the simulation runtime engine and comply the software interface is standard to all objects at that level. Additionally, the interface between the "player", and its subcomponents, "assemblies" and "elements", is also standard. This interface is similar to but NOT exactly like the player to runtime engine interface. Figure 1 represents the J-MASS SSM implementation.

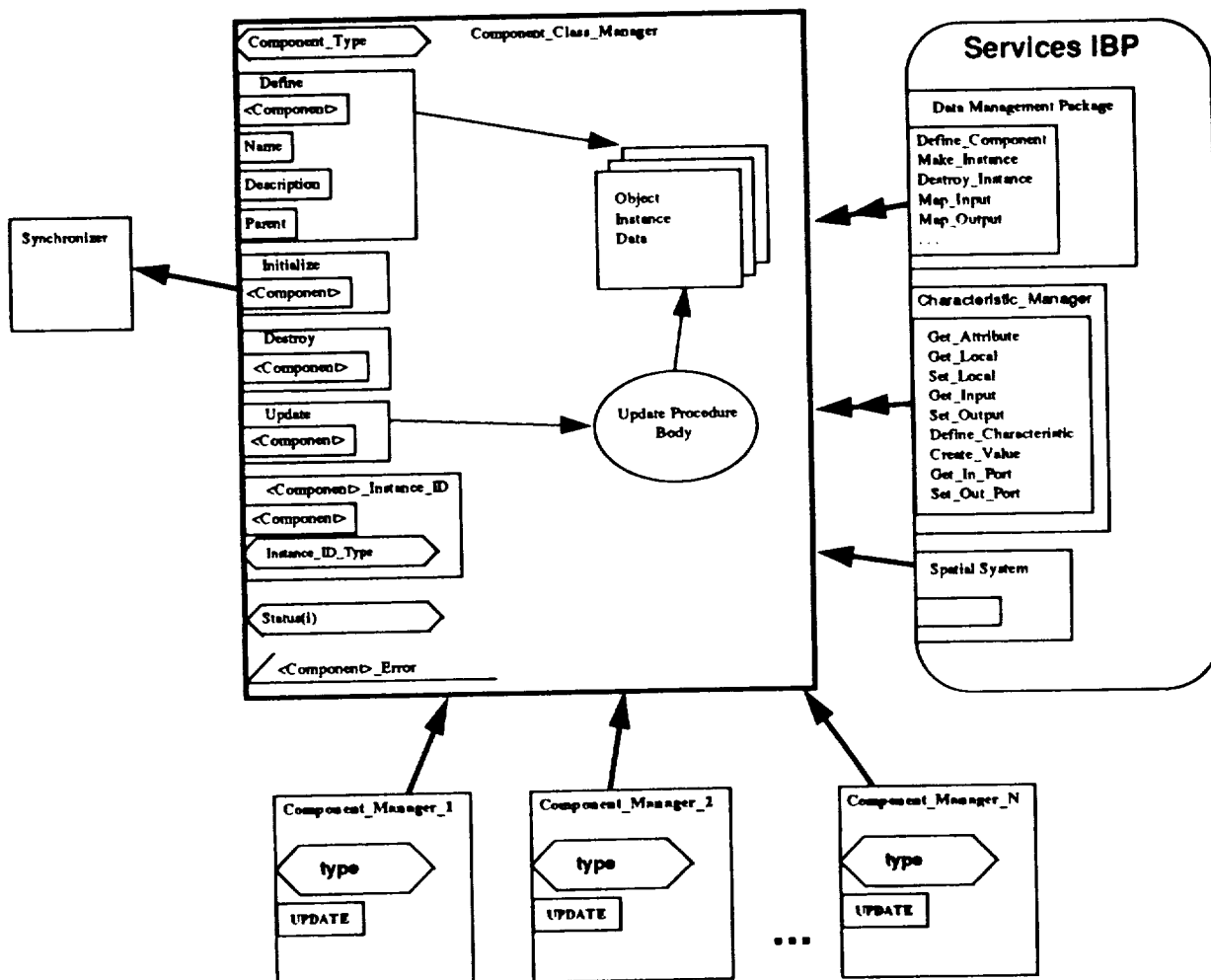
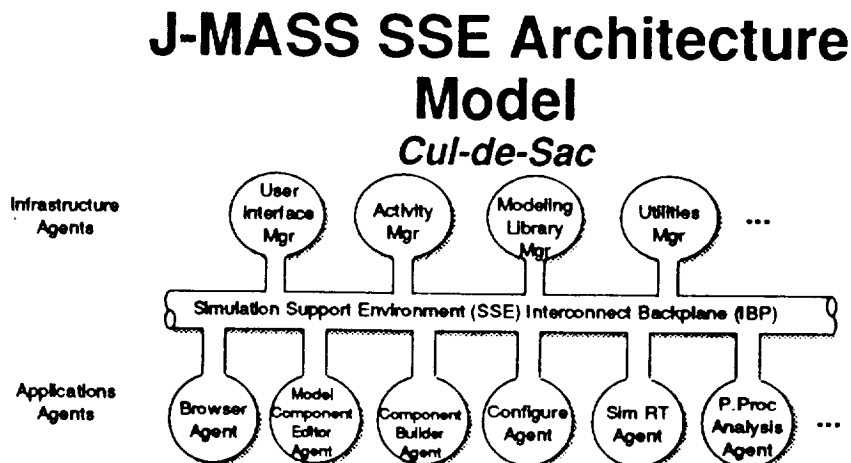


FIGURE 1

OPEN ARCHITECTURE - TOOL INTERCONNECT BACKPLANE

At the Tool Interconnect Backplane (IBP) layer, known in J-MASS as the Simulation Support Environment (SSE) IBP, several backplane methodologies were considered, including the HP Softbench, IEEE P-1175 "Toaster Model", the Atherton Backplane, and, significantly upon the Common Object Request Broker Architecture (CORBA) from the Object Management Group (OMG). In J-MASS terminology, a cul-du-sac model is employed, where each cul-du-sac represents a tool, or potentially a collection of tools or capabilities, referred to as "agents". Each "tool" or "agent" (a software capability), can register as a client/server with the backplane, indicating the service/message traffic of interest. The backplane maintains the knowledge of the other tools that have registered that can either provide the service, or will request the service. This concept is known as message brokering and is powerful for de-coupling the tools from knowledge of other tools on the system. J-MASS has implemented a prototype of its design for this backplane in C on Unix workstations, currently SUN Sparc series, and Silicon graphics. Other platforms in progress include the IBM RS6000, and HP 9000 series, with DEC Alpha, and VAXstations in the plans. Reference Figure 2 for a graphical depiction of this concept.



- **Infrastructure Agents**
 - Agent With Specific Responsibilities
 - Always Installed And Available
- **Application Agents**
 - Connect Directly To SSE Interconnect
 - Register Own Services
 - Request Services To Be Performed
 - Are Loaded / Removed Dynamically
 - Communicate Via SSE Interconnect Message Language Grammar

FIGURE 2

User Modes.

J-MASS has five conceptual "user modes" associated with it. These are "functionally" oriented modes, namely: Develop; Assemble; Configure; Execute; and Post-Process. Each represents a capability that a model developer and/or simulation analyst requires to build, configure, execute, and analyze simulations. Each of these modes can be viewed as an instance of the cul-du-sac methodology. The next series of charts (2 thru 6) depict an instance of the backplane at the tool interconnect layer for each of the J-MASS modes.

Develop Mode/Assemble Mode.

Develop Mode and Assemble Mode provide the model developer with visual mechanisms for constructing model objects/object hierarchies, with data flows represented. Control flows (not currently implemented) will also be depicted so that model developers can separate control/activation of objects from data flow. The graphical information is then translated to ascii "dot" notation, referred to as .DSC (description) files. These .DSC files are then read by an automatic code generator, which generates source code compliant with the Software Structural Model (SSM) in various languages (currently Ada, C++). The SSM is discussed further in the Model Component/Object Standards section. At this point, the algorithms for the lowest level objects in the decomposition must still be described (currently, in the native language thru an editor). The code can then be compiled, linked, loaded and executed. A semantic tool, or "template" editor, is provided to build the semantic "template" information that describes "normal" assembly of the model components, which is done in "Assemble" Mode. Here in develop, the template semantics are generated. See Figures 3 and 4 for a graphical depiction of Develop Mode. Assemble mode permits the connection of the model objects built in Develop Mode visually. The "templates" are populated with actual object instance selections. All of the model components are stored in the modeling library, an object oriented storage mechanism which makes the information about the objects in J-MASS available to all other agents on the backplane.

Application Agents Supporting Develop

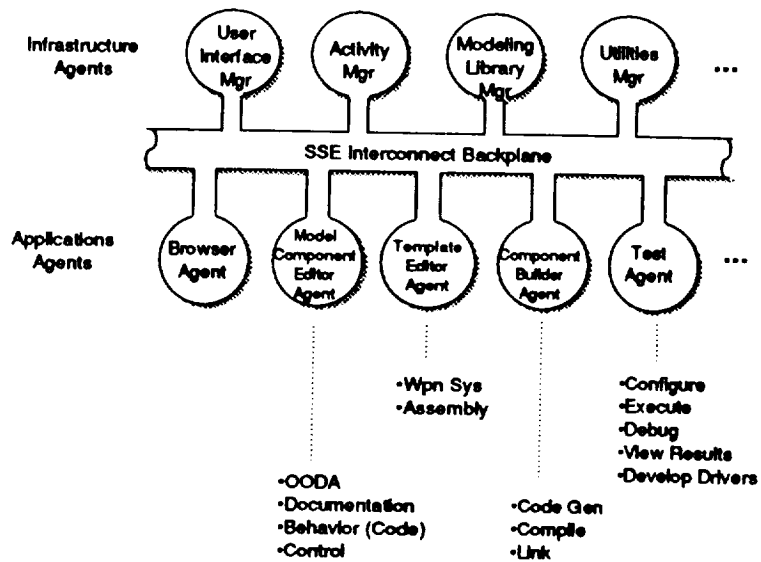


FIGURE 3

Application Agents Supporting Assemble

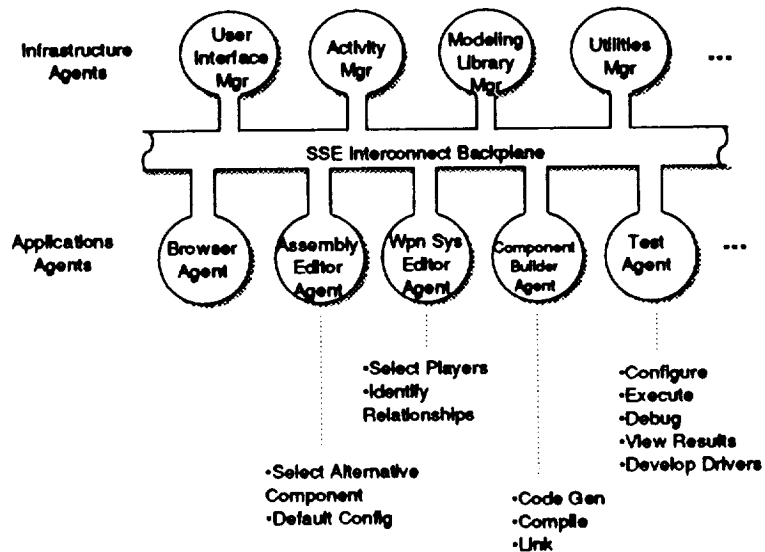


FIGURE 4

Configure Mode.

Configure Mode permits the M&S developer/analyst with the capability to determine simulation characteristics. Model component objects attribute values are populated with values thru a graphical configure tool. Additionally, geographical laydowns, raster maps, etc. are made available to set up the scenarios of the model objects stored in the model library. J-MASS "teams" are formed, whereby player classes are defined, and actual player instances are populated for the teams. This "distribution strategy" is totally configurable by the user. If "legacy" simulations exist, the configure mode will permit the modeler/analyst with the capability to catalogue those models/simulations, and have data passed back in forth sequentially. Eventually, real time synchronized communication between J-MASS compliant and legacy simulations will be achieved. Additionally, if a Distributive Interactive Simulation (DIS) Protocol Data Unit (PDU) generation is desired, the user is able to configure a J-MASS team (collection of players into a single executable). The entire team will then generate PDUs, and the J-MASS spatial system will create "objects" for the incoming DIS entities. The software that provides this capability is the DIS_manager software, and is de-coupled from the standard J-MASS objects, so as not to perturb that interface. Figure 5 depicts the architecture backplane instance for Configure Mode.

Application Agents Supporting Configure

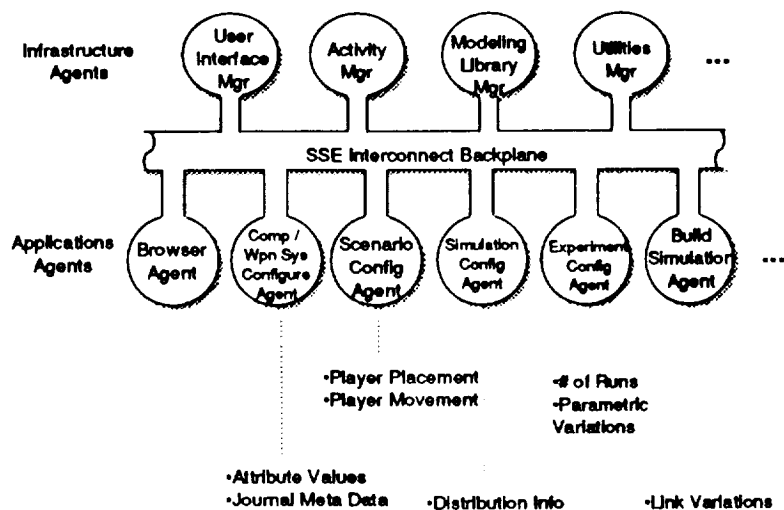


FIGURE 5

Execute Mode.

Execute Mode simply executes the selected simulation. Currently, visualization is accomplished in the Post Process Mode. If the DIS manager software was invoked due to configuration selection, then using "magic carpet" software, the PDUs can be displayed in real time. In work is a real time display of the simulation as it occurs. Figure 6 depicts the architecture instance for the Execute Mode.

Application Agents Supporting Execute

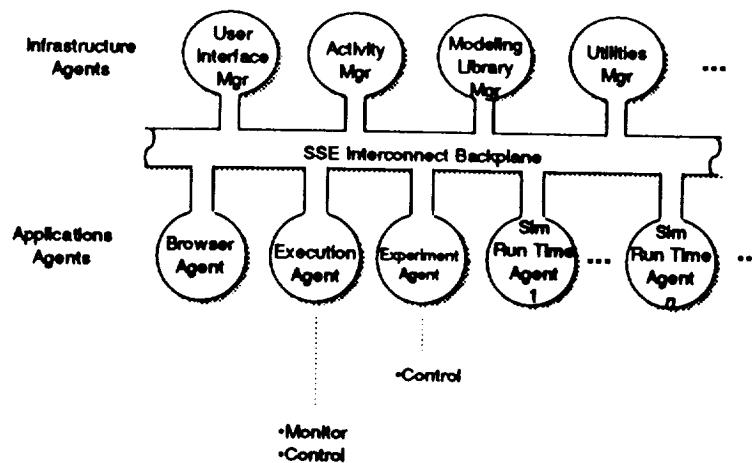


FIGURE 6

Post Process Mode

Post Process Mode is a visualization, both static and dynamic, of the information of interest to the user. This mode includes graphical plotting tools, and animated playback capability. The extraction tool converts the binary journalized data into ascii information. The filter mechanism then prepares it in the appropriate format for the display tool requested. Figure 7 describes the backplane instance for the post process mode of J-MASS.

Application Agents Supporting Post-Process

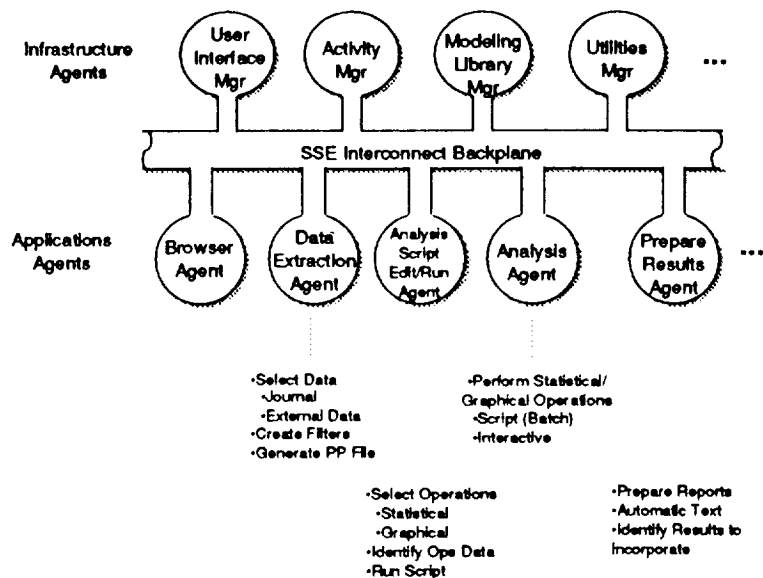


FIGURE 7

SIMULATION RUNTIME AGENT (SRA) ARCHITECTURE

The J-MASS Simulation Runtime Agent (SRA) architecture is depicted in Figure 6. The SRA is "expanded" in this view to show its own architecture. In fact, any agent on the SSE backplane may in fact be another recursive instance of the SSE level. Notice the SRA has its own backplane. The SSE level and SRA level backplanes could in fact be the same. In our current implementation, they are not, but both are distributed in nature using standard Unix (TCP/IP) socket message passing mechanisms. What is important to note in the SRA is the encapsulation of the spatial object, synchronization object, data management object, journalization object, and others away from the model objects. Thus, a true "plug and play" architecture is achieved because any given object may be replaced in the architecture without perturbing the other objects. In the SRA, each team is a single executable using a shared memory implementation, providing significantly faster communication than "inter-team" communication, which uses Unix sockets. Just as the SSE level architecture is distributable, so too are the "teams" within any given SRA. A J-MASS system may in fact have more than one SRA, each communicating over the SSE level backplane. In fact, we plan to demonstrate an Ada SRA with Ada model objects communicating with C++ SRA and C++ model objects over the SSE IBP mechanism. "Players" communicate with each other by placing information on each others "ports" facilities. Players do NOT require apriori knowledge of what team the other player is on, the team synchronizers work with the SRA synchronizer to "locate" the appropriate port. Again, the model objects remain "un-perturbed" with this approach. Journalization of output is accomplished by the journalization object, using state information maintained in the Data Management Package (DMP). In this way, non-intrusive journalizaing occurs. Figure 8 represents the expanded view of the SRA.

J-MASS Architecture Simulation Runtime Agent (SRA) Detail

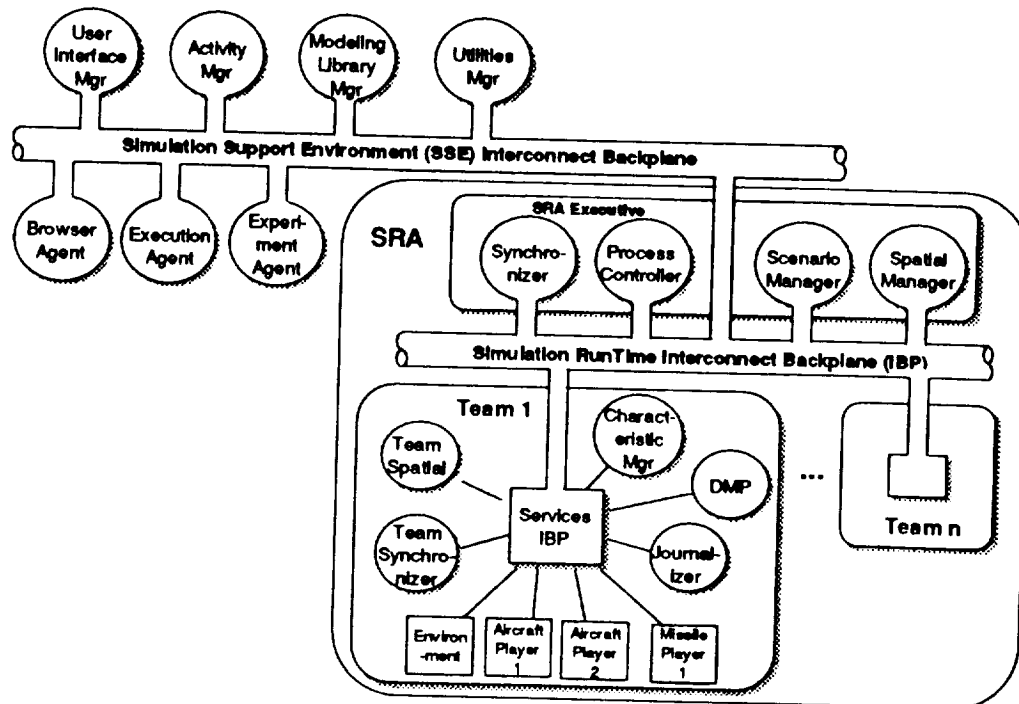


FIGURE 8

COMMERCIAL POTENTIAL

The J-MASS concepts and philosophies are not entirely original. The backplane methodology, message brokering mechanisms have been espoused by OMG and others. However, J-MASS has applied these concepts to a generalized Modeling and Simulation System.

J-MASS brings the idea of standards for digital simulations, both in structure and interface. This guarantees "plug and play" philosophies, both from model components and architecture components point of view. J-MASS espouses the idea of "plug and play" throughout the architecture to include tools, objects (model components), etc.

The J-MASS notion of graphical tool environment coincides with standard commercial technology as well. Expanding that concept which permits (automated) standard compliance with specified standard structures is another potential benefit to the commercial world.

J-MASS itself does NOT prescribe what objects or systems are modelled with its architecture. For example, the object repositories could represent traffic objects, manufacturing objects, weather objects, organizational objects, utility objects, etc. The system is designed so that the M&S communities build object hierarchies and behavior appropriate for the particular domain.

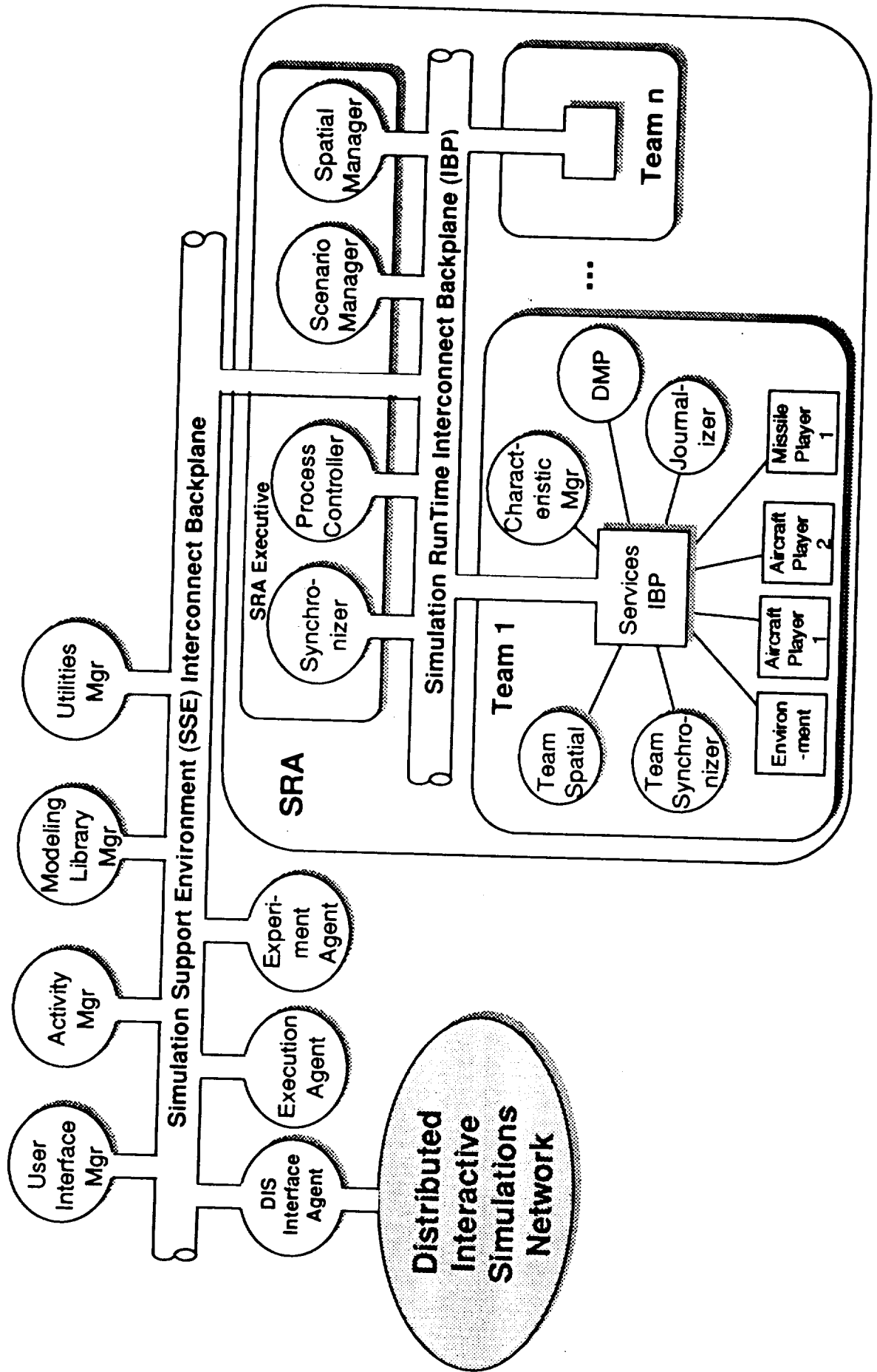
J-MASS / Windows Comparison

J-MASS

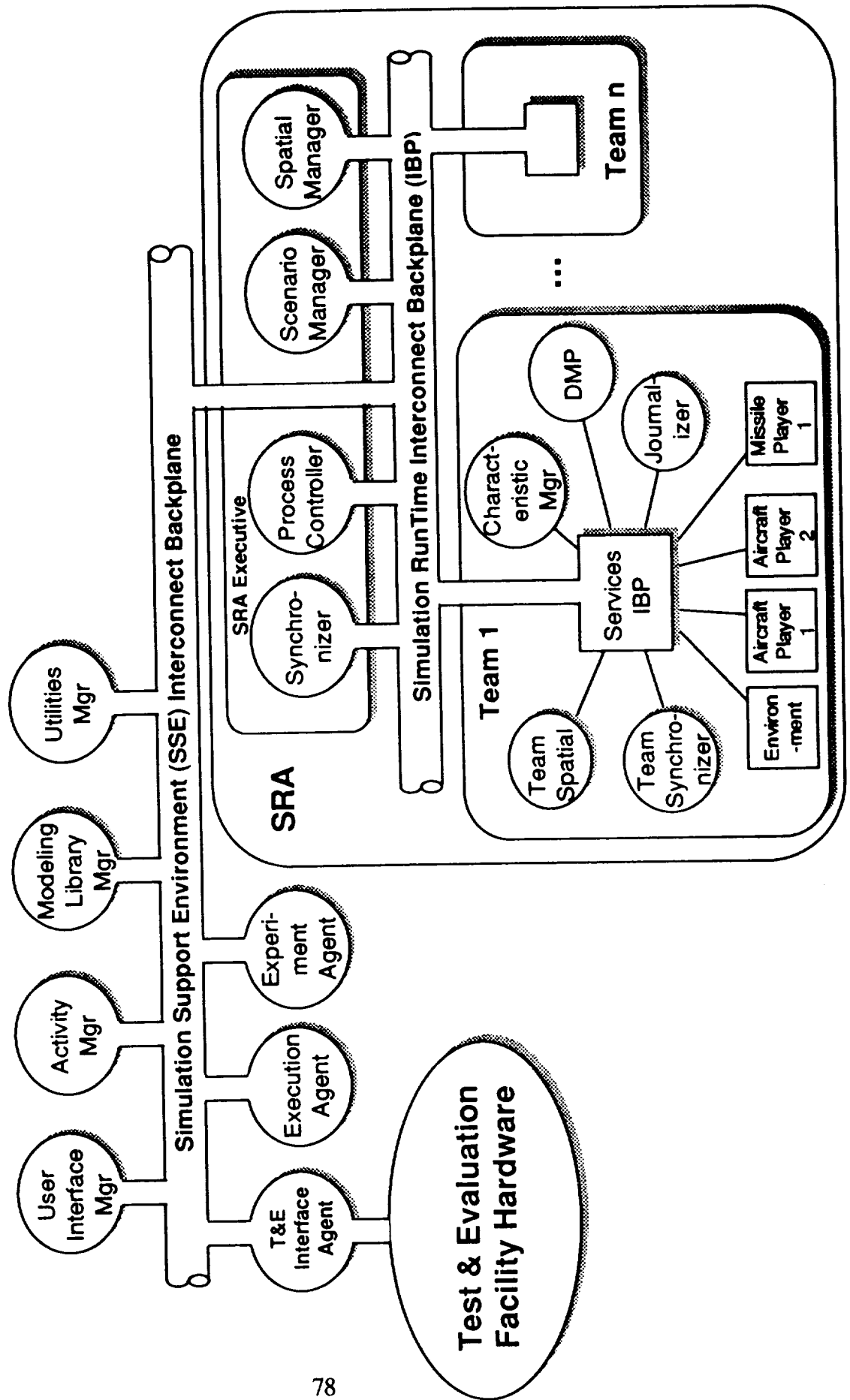
Windows

J-MASS Development Team	Microsoft Corp.
J-MASS Software	Windows Operating System
Tool Interconnect Standard	Windows Interface Standard
User Interface Mgr, Activity Mgr, Modeling Library Mgr, Utilities Mgr	Program Mgr, File Mgr, Print Mgr, etc.
Browse Agent, Test Agent, Simulation Runtime Agent (SRA)	Write, Paintbrush, Terminal, etc.
Specialized Post Processing Tools, Unique SRA, etc.	Word for Windows, Excel, PowerPoint, etc.
Model Interconnect Standard	Word File Format
J-MASS Model Component	Word Document

J-MASS Architecture Extendible to Training

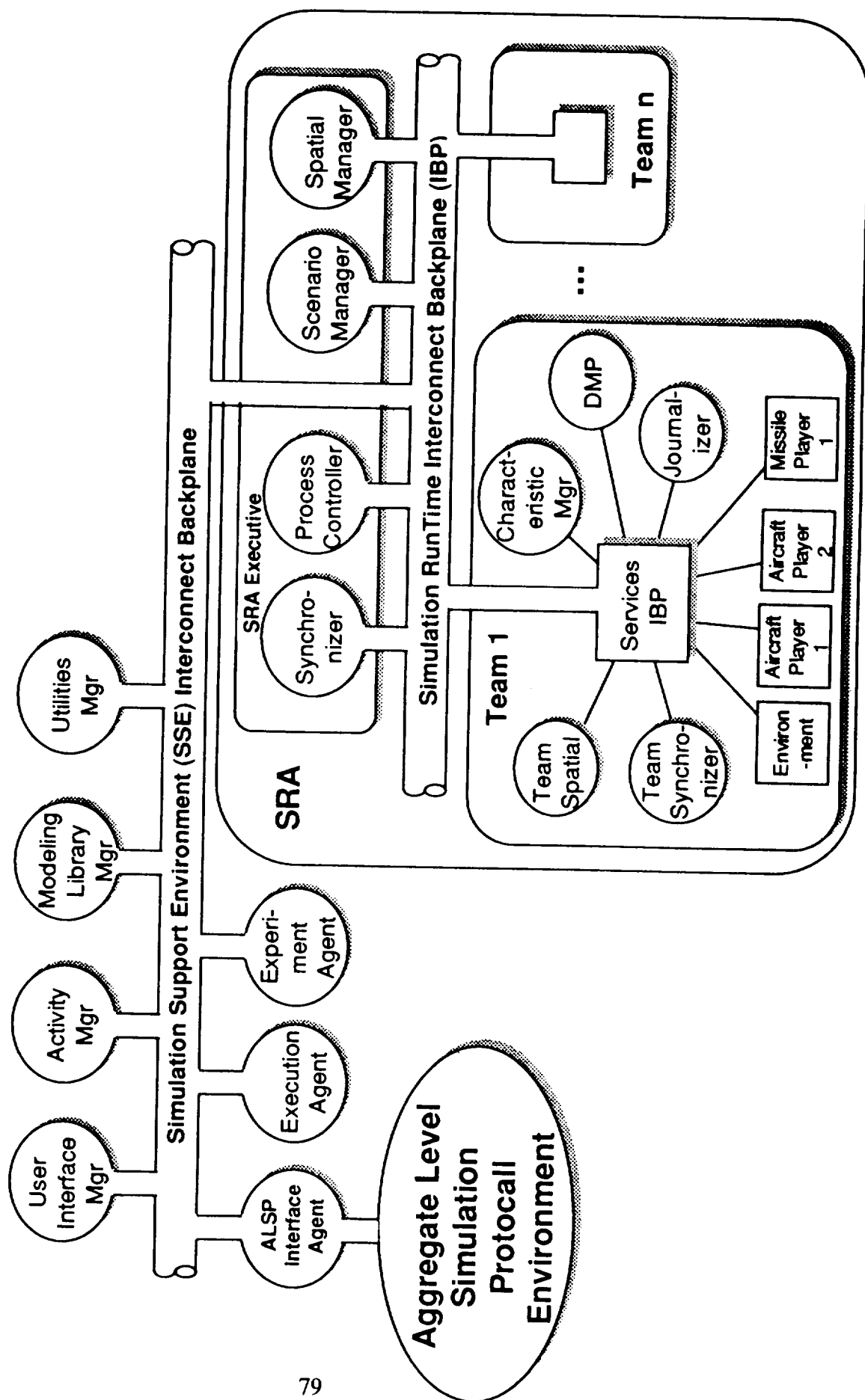


J-MASS Architecture Extendible to Weapon Systems Acquisition



J-MASS Architecture

Extendible to Wargaming



510-61
2493
P- 13

**ANALYTICAL DESIGN PACKAGE - ADP2
A COMPUTER AIDED ENGINEERING TOOL
FOR AIRCRAFT TRANSPARENCY DESIGN**

J.E. Wuerer, PDA Engineering
M. Gran, Wright Laboratory
T.W. Held, University of Dayton

ABSTRACT

The Analytical Design Package (ADP2) is being developed as a part of the Air Force Frameless Transparency Program (FTP). ADP2 is an integrated design tool consisting of existing analysis codes and Computer Aided Engineering (CAE) software. The objective of the ADP2 is to develop and confirm an integrated design methodology for frameless transparencies, related aircraft interfaces, and their corresponding tooling. The application of this methodology will generate high confidence for achieving a qualified part prior to mold fabrication.

ADP2 is a customized integration of analysis codes, CAE software and material databases. The primary CAE integration tool for the ADP2 is P3/PATRAN, a commercial-off-the-shelf (COTS) software tool. The open architecture of P3/PATRAN allows customized installations with different application modules for specific site requirements. Integration of material databases allows the engineer to select a material and those material properties are automatically called into the relevant analysis code. The ADP2 materials database will be composed of four independent schemas: CAE Design, Processing, Testing and Logistics Support.

The design of ADP2 places major emphasis on the seamless integration of CAE and analysis modules with a single intuitive graphical interface. This tool is being designed to serve and be used by an entire project team, i.e., analysts, designers, materials experts and managers. The final version of the software will be delivered to the Air Force in January, 1994. The Analytical Design Package (ADP2) will then be ready for transfer to industry. The package will be capable of a wide range of design and manufacturing applications.

1 INTRODUCTION

ADP2 is an integrated design tool consisting of existing analysis codes and Computer Aided Engineering (CAE) software. The objective of the ADP2 effort is to develop and confirm an integrated design methodology for frameless aircraft transparencies. ADP2 analysis capabilities include: aerodynamic heating, transient thermal response, static and dynamic structure response, optical ray trace, injection molding process simulation, and an aircraft transparency related material properties modeling and databank system. The design process is to be iterative and capable of producing frameless transparency designs, information needed for the design of integral aircraft interfaces, and information needed to support the design of injection molding tooling and the specification of molding process parameters for specific materials.

ADP2 is a second generation analysis system. The initial ADP development was initiated in 1989, References 1, 2 and 3. Since the inception of the original ADP, significant developments in both CAE and design support software relevant to aircraft transparency design have evolved. In addition, certain current design requirements, e.g., optics analysis and material properties modeling, were not addressed in the original ADP. Finally, significant advances in computing hardware have occurred making it possible to perform the required computations on workstation systems as opposed to mainframe platforms.

Recent developments in CAE design tools have introduced the ability to integrate special purpose and commercial-off-the-shelf (COTS) software in a user friendly (intuitive/interactive) environment. That is, to have a single user interface serve the primary analysis functions, specifically:

1. Modeling

- Geometric Modeling (construction and modification)
- CAD and IGES File Import
- Graphics Manipulation

- Meshing
 - Loads and Boundary Conditions Specification
2. Materials Data Management
 - Analysis Code Properties Input
 - Test Data Processing and Reduction
 3. Analysis Module and Module Parameter Specification
 4. Results Evaluation

The design of ADP2 places major emphasis on the seamless integration of CAE and analysis modules with a single intuitive graphical user interface.

2 ADP2 SCOPE

The principal objective of the ADP2 end product is to provide an integrated CAE tool to support the design of frameless (injection molded) aircraft transparencies. This tool is being designed to serve and be used by the entire project team, i.e., analysts, designers, materials experts, and managers. Emphasis is being placed on ease of use with the need for the user to learn only one common graphical user interface.

The ADP2 design places emphasis on the use of standardized methods for electronic communication of data to support the CAE process. Specific examples include: (1) the ability to import CAD and/or IGES files to simplify geometry model development, (2) the ability to store, process and import materials property data to support the design process, and (3) to allow users (including users at different sites) to share analysis results.

Finally, the ADP2 design addresses the issue of cost effectiveness. Where possible, commercial off-the-shelf (COTS) software is specified. This approach alleviates the need for the software owner to support and upgrade the software system as relevant new developments and improvements emerge. In addition, COTS software portability is generally maintained for common engineering work station platforms. Cost effectiveness is also addressed through ease of use and enhanced design team communication. For example, the design team will have the ability to review analysis results from the perspective of their individual needs. Ease of use will allow all members of the design team to communicate through direct use of the ADP2 to support their individual roles.

The specified analysis module requirements for the ADP2 are listed in Table 1, (note the ADP2 column). A comparison of the original ADP and ADP2 analysis module set is shown. It can be seen that the specified upgrades are extensive. In fact, none of the original ADP modules are planned for the ADP2 end product. This is the result of several major developments since the original ADP was defined some four years ago. It is planned that ADP2 will be configured so as to eventually allow the evolution of an ADP based on commercial-off-the-shelf (COTS) software. This potential evolution is indicated as ADP/COTS in Table 1. Although ADP/COTS is not a development target for the present program, it is, however, an important planning issue for ADP supportability and has a direct impact on the ADP2 architecture.

Table 1 ADP Development Phases

MODULE FUNCTION	ADP	ADP2	ADP/COTS
Design Process Manager, Graphical User Interface, and Project Manager	IRIS 4-SIGHT	P3/PATRAN	P3/PATRAN
Modeling/Post-Processing	PATRAN 2.4	P3/PATRAN	P3/PATRAN
CAD Interface	IGES File	P3/UNIGRAPHICS P3/IGES	P3/UNIGRAPHICS P3/IGES
Aeroheating Environment	STAHET	STAHET II	P3/CFD
Thermal Response	TAP	TAP II	P3/THERMAL
Structural Analysis	MAGNA	P3/FEA	P3/FEA

Birdstrike Analysis	MAGNA	X3D	ABAQUS EXPLICIT
Optical Raytrace	NONE	OPTRAN	CODE V
Injection Molding	C-MOLD 1.5	C-MOLD 3.2	C-MOLD 3.2
Materials Properties and Process Data Management Sys.	FTPMAT	M/VISION P3/MATL. SELECTOR	M/VISION P3/MATL. SELECTOR
Results Animation	NONE	P3/ANIMATION	P3/ANIMATION

The primary CAE integration tool for ADP2 and ADP/COTS is P3/PATRAN. The open architecture of P3/PATRAN provides the option to use several solver modules other than the baseline modules indicated in Table 1. A listing of alternative solvers which are commercially integrated with P3/PATRAN are listed in Table 2. This flexibility of P3/PATRAN will expand the opportunity for ADP2 and or ADP/COTS users to use their favorite solvers.

Table 2 Application Modules Commercially Integrated With P3/PATRAN

MODULE FUNCTION	MODULE TRADE NAME
CAD INTERFACE	UNIGRAPHICS CADD5 CATIA Pro/ENGINEER EUCLID-IS IGES PDES/STEP
THERMAL RESPONSE	P3/THERMAL SINDA P3/ADVANCED FEA ABAQUS
STRUCTURAL ANALYSIS	ABAQUS ANSYS MARC MSC/NASTRAN P3/FEA P3/ADVANCED FEA

ADP2 is initially being developed for the Silicon Graphics (SGI) Iris workstation which has a UNIX (IRIX 4.0.5) operating system. A development goal is to have the end product portable to common UNIX based engineering workstations. The specific development workstation is an SGI IRIS CRIMSON / ELAN with 64 megabyte main memory, a 50 megahertz (MIPS R4000) processor, and a 3.8 gigabyte capacity system disk.

The overall architecture of the planned ADP2 end product is shown in Figure 1. The primary CAE integration element of ADP2, P3/PATRAN, provides the basic framework for the executive control to implement application module integration, preprocessing (modeling) capability and post-processing (results display) capability. The basic capability of P3/PATRAN is extended by the ADP2 customization, indicated by the shaded area in Figure 1.

3 ADP2 INTEGRATION

What does the ADP2 Customization consist of? This is a question that many ADP2 users have asked. Comments have been made to the effect that one is not able to determine where P3/PATRAN ends and ADP2 begins. The reason for this is actually one of the advantages of using the customization features of the all new P3/PATRAN as the core of ADP2. Enhancements made to P3/PATRAN via the use of the COTS supported PATRAN Command Language (PCL) tool kit makes all added functionality seamless. In other words, the enhancements actually appear to be a part of P3/PATRAN.

3.1 ADP2 System

The ADP2 system includes several specified commercial and non-commercial software application modules which support the design and manufacturing process. These software products are then integrated with custom software to simplify the work of the engineering analyst.

Specifically, the software modules specified for ADP2 were listed in Table 1. ADP2 System functionally consists of two major parts: the Project Manager and the Application Interfaces. The Project Manager manages all data which the engineer creates and accesses. It also manages the process flow between P3/PATRAN and the various simulation tools. The Application Interfaces integrate the specific simulation tools with P3/PATRAN. The overall architecture of the system was shown in Figure 1.

3.2 Project Manager

The Project Manager is responsible for all items that relate to Project Management and Process Flow. Under most instances, the user will no longer need to keep track of his files. All data within ADP2 is referenced with respect to a specific project or task. The user simply selects the task or project that he is interested in and begins modeling or analyzing. ADP2 will even keep track of where the user was during his last visit to ADP2 and place him at the same task the next time he visits ADP2.

The Project Manager also ensures that the analysis tool requested by the user will be used on the model associated with the specified task. The Project Manager actually informs P3/PATRAN of which analysis tool is to be used with a specific model. So, when the engineer specifies that he would like to perform an analysis, P3/PATRAN will know exactly which analysis to perform.

Another feature of ADP2 is the ability for engineers to share data. An engineer can easily copy data from one project to another. He can also copy data from the outside world (data existing outside of the ADP2 system) into the ADP2 Project Management System.

3.3 Application Interface

The Application Interface consists of all analysis tools as well as their forward and reverse translators to P3/PATRAN. It also consists of the graphical forms interface with which the user interacts with to specify what it is that he would like to do with a specific analysis tool.

For instance, take the case of the application interface for the birdstrike analysis module, X3D. Once the Project Manager has handed the model to P3/PATRAN, the user can use P3/PATRAN to improve upon the model. When the user is ready for analysis, he selects "Analysis" on the menu bar and the X3D graphical forms are displayed to assist the user in specifying the parameters for the type of simulation he would like X3D to perform. These forms are part of the custom code developed for the Application Interface of ADP2. When the user is satisfied with the model and parameters that he has specified, he will "Apply" these parameters and begin the simulation task. In the case of X3D, The model and parameters are automatically input into the forward translator. This translator converts the data into the format required by X3D. This translator is also a part of the custom code developed for the Application Interface of ADP2. After the model and parameters have been translated, X3D is automatically submitted as a background job. Once the X3D job has completed, the results can be imported back into the PATRAN 3 model by use of the reverse translator, another piece of the custom code developed for the Application Interface of ADP2.

4 ADP2 APPLICATION MODULES

There are a total of nine primary modules integrated into the ADP2 analysis tool. An overview of the functionality embedded in each of these modules follows.

4.1 P3/PATRAN

P3/PATRAN is an advanced computer aided engineering (CAE) integration tool introduced by PDA Engineering, Costa Mesa, CA, in June of 1992, Reference 4. P3/PATRAN tool has an open architecture which provides several important capabilities. These include:

1. direct software links to leading CAD systems,
2. direct software links to analysis software programs, both leading commercial and user developed programs,
3. pre- and post-processing capabilities to allow geometry and FEM development, as well as analysis results display,
4. an integrated subset of the materials selection capability from PDA's M/VISION family of products,
5. a fully integrated set of P3/PATRAN analysis modules for performing structural, thermal, computational fluid dynamic, fatigue, and other types of mechanical analyses.

P3/PATRAN has a user friendly "mouse activated" window environment graphical user interface (GUI) which provides execution of the analysis system without the need to remember a particular command syntax. The open system architecture and GUI provide the basic capability to build a user friendly executive control capability for ADP2.

P3/PATRAN has been specified for ADP2 primarily for its capability as a CAE system integration tool. However, there are several analysis modules that are fully developed and completely integrated with P3/PATRAN. These analysis modules include:

1. P3/FEA for comprehensive finite element analysis,
2. P3/ADVANCED FEA which is a non-linear module developed jointly with HKS based on ABAQUS analysis technology,
3. P3/FATIGUE for durability analysis,
4. P3/CFD is a state-of-the-art computation fluid dynamics module for workstations,
5. P3/THERMAL is a state-of-the-art finite element based thermal analysis, and
6. P3/ANIMATION is an advanced, state-of-the-art animation capability co-developed with Intelligent Light.

All of the above modules have specific relevance to the requirements of the ADP. However, the initial development plan includes only two modules. These are P3/FEA which is reviewed in Section 4.5 and P3/ANIMATION which is reviewed in Section 4.7.

4.2 M/VISION

M/VISION is a PDA Engineering materials software system which provides for data visualization, selection and integration, Reference 5. M/VISION uses centralized relational databases organized to reflect the classes of materials information needed in the design-to-manufacturing process:

1. Material: Material source and designations, as well as extension to account for the multiple sources associated with composite materials.

2. Specimen: Detailed information about the specimens including composition and processing specifics.
3. Environment: Specific information about the experimental conditions including temperature and humidity as well as statistics and quality.
4. Properties: Mechanical, physical, thermal, and electrical properties of materials including extensions to account for anisotropy. Curves and rasterized images are represented as well.

Data stored in M/VISION may be directly accessed by P3/PATRAN via the P3/MATERIALS SELECTOR.

M/VISION provides a very significant upgrade for ADP2's materials database management functionality, compared to that in the original ADP. Details of the role of M/VISION in the ADP2 design methodology are presented in Reference 6. The following partial list summarizes several pertinent features provided by the M/VISION COTS product:

1. Complete graphical user interface (GUT) that provides visualization of all materials database functions.
2. Customization of database schema; i.e., attribute list, relationships, etc. Easy to expand database to include more properties and more property types.
3. Units easily changed to SI or any user defined system.
4. Metadata and footnotes provided with data values that enable inclusion of material source, specimen conditions, test environment, test anomalies, etc.
5. Data can be stored and manipulated using an extensive spreadsheet functionality.
6. Database queries using engineering terminology to define search conditions.
7. Full-featured manipulation and storage of tables, graphics, and CAT scan images.
8. Strict adherence to government and industry standards.
9. Provides data importing and exporting features using the following exchange protocols:
 - IGES
 - PDES/STEP
 - PATRAN Neutral File
 - ASCII Text Spreadsheet Files
10. M/VISION databases can be accessed directly or queried using the new P3 "materials selector" functionality (via GUI form driven features).

As a powerful stand-alone product, M/VISION can be utilized in a variety of ways, within the design-to-manufacturing process. Generally, throughout the design process, engineers do not have a central on-line source and electronic access to high quality materials information required for analytical design assessment simulations. Often the materials information that is available is missing critical property values requiring further testing or analytical material synthesis. Therefore, M/VISION's role could be two-fold. As a central source for database management, M/VISION could serve to create, update, and store all the available materials information pertinent to the transparency design community. In this role, M/VISION would be executed and maintained by a Materials and Processes group(s) for the overall design-to-manufacturing process. Materials data, needed by design simulation modules, would be accessed using one of the various transfer protocols listed above.

4.3 STAHET II

STAHET II performs the computations which predict the aerodynamic heating history to the transparency system for a specific aircraft configuration and mission profile. The results of these computations are then used as input boundary conditions to the transient heating analysis (TAP II) for the detailed transparency model, as discussed in Section 4.4.

The STAHET II and TAP II codes are embedded in the STAPAT II (Specific Thermal Analyzer Code for Aircraft Transparencies) development. STAPAT II, rather than a single computer program, is actually an analysis methodology incorporating a set of computer codes. There are 5 major elements of this methodology, specifically:

1. develop the forebody model,
2. compute the aeroheating over the transparency surface using the STAHET computer code and these models (STAHET),
3. develop the transparency finite element model (STABLD),
4. compute the transient, three dimensional transparency system temperatures using the TAP computer code and the finite element model (TAP), and
5. display the models and results using the STAPLT computer code (STAPLT).

Details of the above methodology are reported in References 7, 8, and 9.

The above methodology has several inconsistencies with the ADP2 architecture, specifically with respect to Item 1, forebody model; Item 2, transparency finite element model; and Item 5, display models and results. In ADP2, all of these tasks are handled by P3/PATRAN. In addition STAPAT II was specifically developed for the DEC VAX/VMS computer systems. For ADP2, both STAHET II and TAP II have been ported to UNIX based platforms.

STAHET II represents a significant upgrade to the STAHET code which is embedded in the first generation of ADP (Reference 3). The ground rules for developing the code were that STAHET II must: (1) be user friendly and non-proprietary, (2) provide accurate transparency temperatures without excessive computing resources; (3) use existing methodologies to predict heating rates and temperatures; and (4) retain or improve all original STAHET capabilities, Reference 7.

The general capabilities of the STAHET II code are summarized in Table 3. Details of the indicated capabilities, their assumptions and limitations, and their implementation are presented in References 8 and 9.

Table 3 General Capabilities of STAHET and STAHET II

STAHET	STAHET II IMPROVEMENTS
<ul style="list-style-type: none">• Two Streamline Tracing Methods• Modified Newtonian Pressure Calculation• Laminar and Turbulent Heating Correlations• Boundary Layer Transition Options• Wall Temperature Effect Modeling• Mission Profile (Mach and Altitude) Input for 3-D Geometries• 2-D Wind Tunnel Modeling• Standard Day, Hot Day, Cold Day Atmospheres• Ideal Gas Air Model	<ul style="list-style-type: none">• Simplified Streamline Tracing Techniques Selected as Default• User Input of Streamline Starting Angles Added• Four Shadow Region Pressure Prediction Techniques Added• Wall Temperature Effect Modeling Improved• 3-D Wind Tunnel Model Capability Added• Extension to Hypersonic Flow

The integration of STAHET II into the ADP2 product is limited to those capabilities relevant to current tactical and strategic aircraft. Capabilities relevant to hypersonic vehicle applications, i.e., Mach Number greater than about 4 and altitude greater than 100,000 feet, exist in the basic STAHET II code. The implementation of this

capability would require an extension of the ADP2/STAHET II application interface.

An important feature of the STAHET II integration into ADP2 is the inclusion of (and ready availability of) aircraft forebody geometries. The complete file of 16 STAHET II geometries has been included. Geometries relevant to conventional military aircraft include F-16, F-15, F-4, F-18, and B-1B. Low RCS aircraft, missile, and several hypersonic forebodies also exist in the library. The example of the F-16 forebody configuration is shown in Figure 2.

An example of an ADP2/STAHET II result is provided in Figure 3. Shown is a fringe plot of aerodynamic heating rate over the portion of the forebody relevant to the transparency. The relevant aerodynamic heating parameters are then transferred to TAP II where the predictions of the transient thermal response for the transparency are conducted.

4.4 TAP II

The TAP (Transparency Thermal Analysis) application module performs the transient thermal response analysis for the transparency structure. The primary end result of the analysis is the temperature field as a function of time for the detailed 3-dimensional finite element model. These data may then be queried to examine for critical temperatures and/or be directly handed off to the thermal stress finite element model via the P3/PATRAN FEM Field Interpolator.

The general capabilities of the TAP application module are summarized in Table 4. Details of the above capabilities and their implementation are discussed in References 8 and 9.

An example geometry used to build the TAP finite element model (FEM) is shown in Figure 4. The specific geometry is for the designated "Confirmation Frameless Transparency" (CFT7A) as specified by Wright Laboratories. Example TAP II results are shown in Figure 5. Shown is the temperature field, at a specific mission time point, displayed in the TAP II FEM via the P3/PATRAN post-processing capability.

Table 4 General Capabilities of TAP and TAP II

TAP	TAP II IMPROVEMENTS
<ul style="list-style-type: none"> • 3-D Finite Element Solution • Material Property Data Base • Aeroheating Imposed on External Surface • Thermal Boundary Conditions <ul style="list-style-type: none"> - Generalized Convection - Defog - Anti-Ice (Hot Air Blast) - Electrical Anti-Ice • - Generalized Heat Flux - Radiation-to-Sky - Element-to-Element Radiation • Standard Day, Hot Day, Cold Day Atmospheres • Mission Profile (Mach, Altitude and Time) Input 	<ul style="list-style-type: none"> • Defog Modeling Improved • Cabin Cooling Air Velocity Default Added • Fluid Gap Modeling Improved • External Anti-Icing Improved • Line Source Capability for Convection Added • Earth Radiation Sink Temperature Added • Expanded Material Property Data Base • Viscosity and Molecular Weight Added to Material Property Data Base • Gap Fluids Added to Material Property Data Base • Extension to Hypersonic Flow

4.5 P3/FEA

P3/FEA is a finite element based structural analysis solver that is fully integrated with the P3/PATRAN family of COTS products. P3/FEA has a wide variety of analysis capabilities, solution sequences, element types, and material models, including temperature dependent laminated composite and nonlinear material properties, Reference 10. P3/FEA makes full use of the graphical user interface (GUI) forms system provided by P3/PATRAN. P3/FEA analysis jobs are assembled, submitted, queried, or aborted from within the P3/PATRAN environment. Results evaluations are completely menu driven and fully integrated with the P3/FEA results database.

P3/FEA solves a wide variety of structural problems. An extensive finite element library and set of solution procedures, which can handle very large matrix equations are available in P3/FEA. No translators are needed

to transfer model data from P3/PATRAN to P3/FEA or to transfer results back to P3/PATRAN for post-processing. The close coupling of P3/PATRAN and P3/FEA provides the user with all the efficiencies related to integrated software.

P3/FEA can handle extremely large problems. Models of over one-hundred thousand degrees-of-freedom have been analyzed. A restart capability is provided for large problems that require efficient re-analyses as new cases are defined for existing models.

P3/FEA will provide two basic analytical simulations needed to support the ADP2 transparency design assessments. These are:

1. static analysis (linear and nonlinear) and
2. dynamic normal modes.

The static analysis will include mechanical load environments (pressures) and/or thermal load environments (in-depth thermal gradients). The load/temperature values defined for these analyses are provided by the TAP II aerothermal simulation module solution results. Automated procedures to interpolate the TAP II results onto and into the P3/FEA model are included in the P3/PATRAN FEM field interpolation function. The transparency FEM can include "contact only" nonlinear elements in the regions of attachment to simulate joint details. The nonlinear iterative procedures to establish the discontinuous contact elements' unique combination of contact and gapping is provided as one of the P3/FEA solution procedures. The transparency material properties used in these static analysis simulations can be temperature dependent and/or nonlinear in their defined constitutive relationships. P3/FEA provides the static solution procedures to accommodate temperature dependent element material property evaluations.

The dynamic environments to be simulated by P3/FEA for the transparency design assessments will include normal modes evaluations and possibly subsequent frequency response and random vibration solutions. P3/FEA provides enhanced dynamics analysis capability sufficient to simulate any foreseen transparency dynamic load environment or modal content survey. P3/FEA can solve these problems in either a time or frequency domain.

An example P3/FEA result is shown in Figure 6. Shown is a fringe plot of Von Mises stress on the transparency surface displayed on the P3/FEA finite element model. The illustrated stress is the result of the combined effects of pressure and thermal loading.

4.6 X3D

X3D is an explicit FEA code, used to simulate soft body impact problems, including the birdstrike of aircraft transparencies. It provides substantial improvements over the MAGNA analytical functionality of the original ADP.

The analytical simulation of transparency non-linear transient dynamic response to birdstrike represents a distinctive class of impact structural behavior. The University of Dayton Research Institute (UDRI) has developed and is continuing to enhance a new explicit non-linear dynamic response FEA code, X3D (References 11, 12 and 13). X3D utilizes an explicit solution approach similar to other well-known FEA codes; e.g., DYNA3D, WHAMS, and ABAQUS Explicit. These solution methods and codes have been widely used for the numerical simulation of a variety of shock and wave propagation problems. Impact problems in general can be dominated by complicated contact surface conditions which can require very small numerical integration time steps which in turn remove the solution advantages of an implicit solution approach.

X3D provides functionality to idealize both the bird (impactor) and the transparency (target) and to allow them to come into transient contact with the physics of momentum transfer defined by the non-linear explicit solution algorithm solution results. That is, no analysis assumptions regarding contact pressure time histories or spatial distributions are required.

X3D has been evaluated through executions and correlations with benchmark test cases; i.e., impacting Taylor cylinder, exploding cylindrical shell, and F-16 centerline transparency impact. Although documented validation problems are limited, those completed to-date show promising correlations and demonstrate the improved simulations possible with the use of X3D code.

The X3D impact dynamics code provides a significant numerical tool kit to simulate the complex soft body impact environment of transparency birdstrike:

1. FEA Elements

- Both 3D solid (HEX and TET) and 2D layered shell elements are provided. The 2D layered shell element accommodates soft interlayers (plane selections do not remain plane over the overall layered shell thickness)

2. Material Models (2D)

- Elastic-Plastic, Rate Sensitive, Isotropic
- Linear Elastic, Orthotropic, Brittle Failure
- Viscoelastic with failure

3. Material Models (3D)

- Elastic-Plastic, Rate Sensitive, Isotropic
- Same as above with Discontinuous P-V
- Newtonian Viscous Fluid

4. Automated Contact Surface Evaluation Procedures

- Slave and Master node sets that are nonlinearly evaluated for contact

5. Simple "Rigid Wall" designation procedures

6. Linked element lists for failure assessments

7. Element and integration stabilization features

An example X3D result is shown in Figure 7. The deformed transparency and bird finite element models are displayed, for the specified time point. A fringe plot of Von Mises stress is displayed on the model.

4.7 OPTRAN

OPTRAN is a raytrace code which evaluates the optical quality of aircraft transparencies subjected to operational load conditions. The code was developed by the University of Dayton Research Institute, References 14 and 15.

The raytrace optical code is interfaced to finite element thermal and stress analysis codes to permit the effects of operational loads to be modeled. Thermal, displacement, and stress field definition data computed by the finite element codes are input to the optics module. This information is required to compute the orthotropic indices of refraction throughout the material volume of the aircraft transparency. This computation is performed at each step along the propagation path of each ray.

The optics code tracks rays of various wavelengths through the transparency. The deformed geometry generated by the stress analysis is used to determine angles of reflection and refraction at transparency layer boundaries. Birefringent indices of refraction are computed as a function of material, temperature and stress state at the refracting surfaces and within the transparency material.

Key results include angular deviation, transmittance, and polarization effects over specified regions of the transparency. Displacement vectors and deformed grids can also be generated.

4.8 C-MOLD

The C-MOLD product is a family of computer codes designed to support the design of tooling and specification of process parameters for fabrication by injection molding, Reference 16. The modular product supports both the processing of thermoplastic and reactive materials. C-MOLD is a commercial-off-the-shelf (COTS) product developed and marketed by AC Technology, Ithaca, NY. AC Technology also provides materials characterization services and maintains a materials database relevant to the injection molding of plastics.

The frameless transparency development is concerned with the use of thermoplastic materials. Therefore, there are three primary processes requiring simulation. These are:

1. the process of the initial filling of the mold (C-FLOW),
2. the post-filling process where shrinkage occurs (C-PACK), and
3. the transient cooling of the part prior to mold separation (C-COOL).

The product modules which address these processes are defined in parentheses above.

The C-FLOW analysis models the mold filling process as a generalized Hele-Shaw (very slow motion) flow (Reference 16). The flow conditions are for an incompressible viscous polymeric melt under non-isothermal conditions and symmetric thermal boundary conditions. The numerical solution is based on a hybrid finite-element/finite-difference method to solve pressure and temperature fields, and a control-volume method to track moving melt fronts. Details of the analysis methodology are presented in References 17 and 18.

The C-PACK analysis module extends the above analyses module to include the effects of asymmetric thermal boundary conditions. A set of unified governing equations for the flowfield is used throughout the filling and post-filling stages. The analyses can model a three-dimensional, thin cavity with a melt-delivery system that may contain cold or hot, circular or non-circular runners. The influence of shrinkage is also included. Details relevant to the C-PACK analysis are presented in References 19, 20, and 21.

C-COOL is a three-dimensional mold cooling simulation to assist in designing the cooling channel system for plastics injection molding processes. The capability exists to model a homogeneous, three-dimensional mold with a thin cavity and with a cooling system that contains circular or non-circular channels, baffles and bubblebs. A channel network analysis within the program within C-COOL predicts flow rates in different cooling lines.

The C-COOL module uses a strategy which minimizes input data requirements, user time and computer memory requirements. Heat transfer within the polymer melt is treated as transient, local, one-dimensional heat conduction with static solidification. Heat transfer within the mold is treated as transient, three-dimensional conduction. Heat exchange between the channel surfaces and the cooling fluid is treated as steady and is accounted for using correlations for the convective heat transfer coefficient. To solve the relevant governing equations simultaneously, C-COOL uses a hybrid scheme consisting of a modified, three-dimensional boundary element for the mold region and a finite difference method with a variable mesh for the melt region. These two analyses are coupled iteratively to match the temperature and heat flux at the mold/melt interface. A special algorithm has been developed which reduces the computational memory requirements by a factor of 100, compared to the requirements for a traditional approach.

An example C-MOLD result is shown in Figure 8. Shown is a result from the C-FLOW module, which shows the melt front as a function of time during the mold filling process. The example modeled the Configuration Frameless Transparency, using a representative polycarbonate resin.

4.9 P3/ANIMATION

P3/ANIMATION is a powerful tool that offers interactive visualization, animation, photo-realistic rendering and video tape output of geometry and results data, Reference 22. It is designed to assist engineers in the investigation and presentation of data which are normally very difficult to visualize. With P3/ANIMATION, information can be displayed in a number of different ways including: Wireframe, Hidden Line, Solid Shaded and Fringed models. Display can be further enhanced with the use of motion to view the data from a variety of angles and a host of other variables such as transparency, surface coloring and shading characteristics.

P3/ANIMATION uses on-screen animation to show geometry and results as dynamic moving pictures. After completing an analysis in P3/PATRAN, static, modal and transient data sets are read directly into

P3/ANIMATION. The user then has complete control over visualization in both space and time. Models may be positioned or rotated to gain a better view and any single time step or range of time may be examined. Fringe plots and arrow fields are used to display scalar and vector data, respectively, and model deformations may also be displayed. Sophisticated timing controls can be used to zoom in on portions of the analysis that are particularly interesting, while skipping over portions of less interest.

The Animator Tool can be used to create simple or key frame animations. In simple animations, models cycle through the results data while rotating about a single axis. In key frame animations, different groups can be posed with various rotations, scales and rendering transforms applied. Posing allows for much more creative and illustrative animations, in which many attributes and transforms can be set or varied over time.

Once an animation is defined, the Flipbook Tool can be used to compute sequences of frames or flipbooks for subsequent playback on any X Windows device in the network. Flipbook images are created and displayed using available graphics hardware to increase performance.

5 CONCLUDING REMARKS

ADP2 has been designed specifically to support the Air Force Frameless Transparency Program. But the package will be capable of a wide range of design and manufacturing applications. The completion of the planned work will provide an important suite of tools to aid in the design and performance evaluation of injection molded transparency systems. The validation of these tools will be a critical aspect of the ADP2 development process.

ADP2 provides a fully integrated methodology relevant to the general problem of aircraft transparency design. A key aspect of ADP2 is the seamless integration of the analysis modules with P3/PATRAN, an advanced commercially supported CAE integration tool. This integration provides a single form driven user interface which serves to manage the analysis process and the analysis module input and output (results) data. The user remains in an intuitive environment and is freed of the complexities and/or peculiarities of the computer operating system throughout the entire design process. The support provided by the employment of a commercial CAE integration tool and the selection of state-of-the-art application modules will provide Air Force and its contractors with a cost effective tool with a significant life potential.

6 REFERENCES

1. Turner, D.L., "Interim Report Draft III: Development Plan for Frameless Transparency Analytical Design Package," Contract No. F33657-84-C-0247, CDRL No. 34002, Report No. FZM-719-001, General Dynamics Corp., Ft. Worth, TX; 28 June 1990.
2. Hunten, K.A., "Analytical Design Package Development Report," Contract No. F33657-84-C-0147, CDRL No. 34002, Report No. FZM-719-009, General Dynamics Corp., Ft. Worth, TX; 21 November 1991.
3. Hunten, K.A., Analytical Design Package Version 1.2c Operation Manual, Contract No. F33657-84-C-0247, CDRL No. 34002, Report No. FZM-719-010, General Dynamics Corp., Ft. Worth, TX; 22 November 1991.
4. P3/PATRAN User Manual, Vols. 1, 2, 3, 4, Publication No. 903000, PDA Engineering, Costa Mesa, CA; September 1992.
5. M/VISION, Materials Visualization, Selection and Data Integration, User Manual, Publication No. 2190011, PDA Engineering, Costa Mesa, CA; September 1990.
6. Mack, T.E., Kipp, T.E., Whitney, T.J., and Gran, M., "The Use of Computerized Materials Data in ADP2," Conference on Aerospace Transparent Materials and Enclosures. (Sixteenth), U.S. Air Force Wright Laboratories; 9-13 August 1993.
7. Varner, M.O., et al, "Specific Thermal Analyzer Program for High Temperature Resistant Transparencies for High-Speed Aircraft (STAPAT)," Report No. AFWAL-OTR-84-3086, Volumes I, II and III, AD B089 497, AD B1090 894L, and AD B090 896L, Wright Research and Development Center, Wright-Patterson AFB, OH; October 1984.

8. Bowman, B.L., "Hypersonic Thermal Analysis for Aircraft Transparencies, Vol. I: STAPAT II Description," Report No. WRDC-TR-90-3053, Wright Research and Development Center, Wright-Patterson AFB, OH; September 1990.
9. Bowman, B.L., "Hypersonic Thermal Analysis for Aircraft Transparencies, Vol. II: STAPAT II User's Manual," Report No. WRDC-TR-90-3053, Wright Research and Development Center, Wright-Patterson AFB, OH; September 1990.
10. P3/FEA Application Module User Manual, Publication No. 903006, PDA Engineering, Costa Mesa, CA; September 1992.
11. Brockman, R.A. and Held, T.W., "Explicit Finite Element Method for Transparency Impact Analysis," Report No. WL-TR-91-3006, Wright Laboratory, Wright-Patterson AFB, OH; June 1991.
12. Brockman, R.A. and Held, T.W., "X3D--3D Explicit Finite Element Analysis--Tutorial," University of Dayton Research Institute, Dayton, OH; July 1991.
13. Brockman, R.A. and Held, T.W., "X3D Users' Manual, Updated for X3D Version 3.04," Report No. UDR-TR-92-59, University of Dayton Research Institute, Dayton, OH; April 1992.
14. Fielman, J.W. and Loomis, J.S., "Optical Analysis of Aircraft Transparencies (OPTRAN), Volume II: Theoretical Manual," Report No. WRDC-TR-90-3058, Volume I, Wright-Research and Development Center, Wright-Patterson AFB, OH; October 1990.
15. Fielman, J.W. and Loomis, J.S., "Optical Analysis of Aircraft Transparencies (OPTRAN), Volume I: OPTRAN User's Manual," Report No. WRDC-TR-90-3058, Volume I, Wright-Research and Development Center, Wright-Patterson AFB, OH; June 1990.
16. CMOLD Reference Manual, AC Technology, Ithaca, NY; 1992.
17. Hieber, C.A. and Shen, S.F., "A Finite Element/Finite-Difference Simulation of the Injection Molding Filling Process, *Journal of Non-Newtonian Fluid Mechanics*, Vol. 7; 1980.
18. Wang, V.W., Hieber, C.A., and Wang, K.K., "Dynamics Simulation and Graphics for the Injection Molding of Three-Dimensional Thin Parts," *Journal of Polymer Engineering*, Vol. 7, No. 1; 1986.
19. Chiang, H.H., et al, "Integrated Simulation of Fluid Flow and Heat Transfer in Injection Molding for the Prediction of Shrinkage and Warpage," Vol. ASME-HTD-175/MD-25, pp. 133-146; 1991.
20. Chiang, H.H., Hieber, C.A., and Wang, K.K., "A Unified Simulation of the Filling and Postfilling Stages in Injection Molding. Part 1: Formulation and Part 2: Experimental Verification," *Polym. Eng. Sci.*, Vol. 31, pp 116-139; 1991.
21. Hieber, C.A., Chapter 1 in *Injection and Compression Molding Fundamentals*, A.I. Isayev, Ed., Marcel Dekker, New York; 1987.
22. P3/ANIMATION Application Module User Manual, Publication No. 903003, PDA Engineering, Costa Mesa, CA, February 1993.

Assembly Flow Simulation of A Radar

W. C. Rutherford, P. M. Biggs
AlliedSignal Inc., Kansas City Division*
Kansas City, MO 64141

2434
P-4

Abstract

A discrete event simulation model has been developed to predict the assembly flow time of a new radar product. The simulation was the key tool employed to identify flow constraints. The radar, production facility, and equipment complement were designed, arranged, and selected to provide the most manufacturable assembly possible. A goal was to reduce the assembly and testing cycle time from twenty-six weeks to six weeks. A computer software simulation package (SLAM II) was utilized as the foundation for simulating the assembly flow time. FORTRAN subroutines were incorporated into the software to deal with unique flow circumstances that were not accommodated by the software. Detailed information relating to the assembly operations was provided by a team selected from the engineering, manufacturing management, inspection, and production assembly staff. The simulation verified that it would be possible to achieve the cycle time goal of six weeks. Equipment and manpower constraints were identified during the simulation process and adjusted as required to achieve the flow with a given monthly production requirement. The simulation is being maintained as a planning tool to be used to identify constraints in the event that monthly output is increased. "What-if" studies have been conducted to identify the cost of reducing constraints caused by increases in output requirement.

Introduction

In 1989, designers at Sandia National Laboratory/New Mexico began the process of designing a new radar. The radar sub-assembly includes nine hybrid modules and a printed wiring board interconnected with .047-inch coaxial cables and two flat flexible cables. Seven of the hybrid circuits modules were designed to be mounted on an aluminum plate that is connected to a printed wiring board with a flat flexible cable. Two hybrid circuits are attached to the printed wiring board. The mounting plate and printed wiring board are folded together, fastened in place, then mounted into an outer housing. The design is the product of a joint effort between the Sandia designers and AlliedSignal manufacturing engineers. The goal was to create a design that was consistent with the manufacturing capabilities at the Kansas City Division. The modular design was selected so that hybrid modules could be built individually then assembled onto a mounting plate. Previous radar products had housed hybrid assemblies in multi-cavity housings with difficult interconnections and an extreme assembly environment. A hybrid failure and subsequent rework would necessitate that good product be subjected to the rework environment. The modular design allows for "drop-in" replacement of hybrid modules.

* Operated for the U. S. Department of Energy under contract No. DE-AC04-76-DP00613
Copyright AlliedSignal Inc., 1993

Past radars had cycle times as high as twenty-six weeks from the receipt of electrical and mechanical components to the completion of an assembly. Typically, a design change would effect numerous assemblies because of the work-in-progress (WIP). In order to become more responsive to design changes, the goal was to produce a radar in twelve weeks at the onset of production and reduce that cycle time to six weeks early in production. A dedicated manufacturing department was built. It contained all the manufacturing and test equipment except equipment for laser marking, radiography, potting, and welding. The equipment and workstations were arranged to economize on the distance traveled between processes. Some compromises were needed since the tester sizes and environment physically limited their location. When compromises were made, emphasis was placed on keeping the lines of communication between assemblers and inspectors open.

Simulation Model Preparation

Having designed the radar for manufacturability, the next steps in the journey to produce a new radar with a 75% reduction in cycle time were 1) verify that the goal was achievable, and 2) implement those controls within the manufacturing area that would insure efficient flow through the assembly process.

SLAM II and TESS, software packages licensed from Pritsker Corporation, were available to be used to conduct a computer simulation of the product flow from the receipt of component parts to completion of the radar. The SLAM II software along with additional FORTRAN code added to handle special cases was necessary for the final model to be designed for specific product flow of the radar.

The initial attempt for the simulation model was based on process information gathered through conversation with the engineers and electronic assemblers. The assumptions used for the early simulation were:

- The radar flow would be a "pull" controlled system.
- Lot size = 1.
- Sub-assemblies between functional test were minimized.
- Limiting equipment resources were accounted for.
- Times were estimates from conversations with assemblers and engineers.
- Testers were available for two shifts.
- Personnel was assumed to be available on demand.

Based on the above assumptions, the cycle time realized was twenty-nine days (six weeks) with a monthly output of seven radars.

Formation of Cross-functional team

The desired cycle time was verified with the early model but the output of seven radars per month was short of the eleven radars per month required. It was obvious that the simulation was a useful tool but the data used to operate the model would need to be more precise. It became apparent that a cross-functional team was needed to insure that the model would precisely represent the assembly processes. AlliedSignal had recently committed to using a Total Quality (TQ) nine step problem solving model. The approach appeared to be a perfect match for building the discrete event simulation model.

A team was formed that included personnel with the following job descriptions:

Process engineer	Inspector
Electrical product engineer	Sandia design engineer
Simulation engineer	Quality engineering
Industrial engineer	Electronic assemblers
Planner	Senior project engineer
Mechanical product engineer	

The team members were selected based on their knowledge of the radar function and processes required for assembly and testing. The TQ approach was presented to the team in four eight-hour-day training sessions. The team focused on "How can the the radar cycle time be reduced to six weeks or less?" The discrete event simulation program would be the tool to verify the cycle time improvements.

Refining the Process Flow

The team began the process of analyzing the current process for producing the radar. The process of collecting assembly steps, times, equipment capabilities, assembler classifications, and inspection points began. The assembly steps were identified in detail and written in the form of a flow diagram. The manufacturing and test equipment was evaluated to establish the most efficient utilization. In some cases, like parts were processed together in batch equipment, temperature cycles were commenced at specific times each shift and some testers were identified for multi-shift utilization. Technical factors such as process schedules for solder reflow of thick film and thin film hybrid network technology impacted the flow sequence decisions. Times were attached to each of the steps. The times were estimates based on experience gained manufacturing prototype parts.

The final radar assembly includes ten sub-assemblies that require electrical testing or circuit tuning. In the case of the hybrid modules two test are required, one before lids are installed and one after lid installation. A system of buffers were established in the model that would signal the start of a sub-assembly once a sub-assembly in work passed its electrical function test. Ideally it is desirable that one sub-assembly is built, completely tested, and qualified before another sub-assembly is started. A goal is to minimize the number of sub-assemblies requiring rework if a process or component causes electrical failures. To improve the output it was necessary to identify interim buffers for hybrid assemblies. The team agreed that the electrical test performed before a lid is attached to a hybrid typically identifies defective product, therefore, a buffer could be added at that point. This is an example of the compromises needed to reduce the cycle time and achieve the output required. The team established the following revisions to the assumptions list:

- Lot size = 1.
- All resources needed as required.
- One shift for assembly.
- Two shifts for testing and tuning.
- Infrared vacuum soldering restricted by part type.
- Two final testers.
- One tester for HMC-I,O ,R, and V.
- One tester for HMC-F, M, T.
- One tester for HMC-L.
- One tester for HMC-C/PWA assembly.
- One burn-in tester for the channel assembly.

One tuning system.

Temperature cycles begin on second shift (eighteen hour cycle).

Vacuum bake starts at beginning of first shift.

This addition of detail and adjustment of equipment utilization resulted in a simulation cycle time of four weeks with 16.7 radars/month output.

The simulation is being refined further to include restrictions to personnel. Classifications and skills of personnel are being included in the model. It is expected that limited personnel will have some impact on the cycle time and monthly output. It is also expected that the simulation can assist in determining the correct mix of skills and classifications of personnel necessary to meet the schedule and cycle time goals.

"What-if"

Soon after the team had developed and ran the simulation model, the impact of doubling the monthly output was considered. The engineers involved evaluated the equipment and tester utilization and presented a list of estimated additional requirements. The simulation engineer was asked if he could factor this into the model and create an equipment utilization prediction. This task was undertaken and eighteen hours later a list of additional equipment and tester requirements was presented along with analysis of the those items that did not need duplication but were highly utilized. The utilization analysis proved especially useful since the potential for flow constraints was more visible.

Next Step

A new team, that includes some members of the simulation team, has been formed to implement the controls on the factory floor that were established for the simulation model. A discipline will be required to insure that part flow is a first-in-first-out pattern and parts are not allowed to be placed in work when the WIP is at its maximum limit. It is expected that modifications to the process and the simulation model will be made to insure that the cycle requirements and output is achieved. It is also expected that the simulation model will be maintained as a tool to evaluate "what-ifs" driven by schedule changes.

Conclusion

The combination of the cross-functional team and computer simulation model created an early need for detailed understanding of the radar assembly processes required to build a radar, created an ownership attitude among the radar team members, provided a tool for future analysis of cycle times and radar output, and established a foundation for an implementation team that will actually produce the radar.

CONFIG - INTEGRATED ENGINEERING OF SYSTEMS AND THEIR OPERATION

Jane T. Malin
Automation & Robotics Division, Engineering
NASA Johnson Space Center
Houston, TX 77058

Dan Ryan
MITRE
Houston, TX 77058

Land Fleming
Lockheed
Houston, TX 77058

5/2-61
2495
p. 8

ABSTRACT

This article discusses CONFIG 3, a prototype software tool that supports integrated conceptual design evaluation from early in the product life cycle, by supporting isolated or integrated modeling, simulation, and analysis of the function, structure, behavior, failures and operation of system designs. Integration and reuse of models is supported in an object-oriented environment providing capabilities for graph analysis and discrete event simulation. CONFIG supports integration among diverse modeling approaches (component view, configuration or flow path view, and procedure view) and diverse simulation and analysis approaches. CONFIG is designed to support integrated engineering in diverse design domains, including mechanical and electro-mechanical systems, distributed computer systems, and chemical processing and transport systems.

INTRODUCTION

The core of engineering design and evaluation focuses on analysis of physical design. Today's computer-Aided Engineering (CAE) and Product Data Management (PDM) software packages can support concurrent engineering, bringing together engineering and production design. However, depending on the engineering domain, they are oriented toward geometry or continuous process and control parameters. They do not provide enough support for conceptual design early in the life cycle or for engineering for operation, fault management, or supportability (reliability and maintainability). Integrated modeling and analysis of system function, structure, behavior, failures and operation is needed, early in the life cycle. The same models can also be reused to support design of fault management software and procedures for the system.

Benefits of concurrent engineering include reduced costs and shortened time for system development. Benefits of engineering for operations and supportability include more robust systems that meet customer needs better and that are easier to operate, maintain and repair. Benefits of reuse in the design of software and procedures include faster software development and more robust fault management.

Conventional system modeling approaches were not designed for evaluating conceptual designs early in the system life cycle. These modeling approaches require more knowledge of geometric or performance parameters than is usually available early in design. Thus, designers rely on "engineering judgment" or systems engineering analysis for early design evaluation. Usually, there is not a traceable path from these analyses to the conventional simulations that are done later. Also, these conventional simulations are often too special-purpose to support evaluations of operability, diagnosability, and supportability.

A more abstracted level of modeling would be sufficient for early conceptual design definition and evaluation, and would also remain useful for some later analyses. Component-connection models are one such useful abstraction. Discrete event models are another useful abstraction. Discrete event simulation technology combines both abstractions, and has been used extensively for evaluation of conceptual designs of equipment configurations in operations research (3). In CONFIG, these abstractions, with some enhancements, are also proving to be useful in defining and evaluating conceptual designs for several types of systems.

Design Goals of CONFIG

When the CONFIG project began, the goal was to support simulation studies for design of automated diagnostic software for new life-support systems (9). The problem was to design an "expert system" on-line troubleshooter before there was enough experience with the system for there to be an expert. The design engineer could use a model of the system to support what-if analyses concerning how failures would propagate, interact, and become observable and testable. With these analyses, development of detection and diagnosis procedures could proceed in parallel with system design. This activity is similar to Failure Modes and Effects Analysis (5), but with comparative simulations of failure effects for the purpose of developing diagnostic software. Conventional simulation software was not up to this challenge, but discrete event simulation software, with enhancements, seemed promising. CONFIG supports the use of qualitative models for applying discrete event simulation to continuous systems, and the use of graph analysis on component-connection models.

CONFIG is designed to model many types of systems in which discrete and continuous processes occur. The CONFIG 2 prototype was used to model and analyze effects of failures in: 1) a simple two-phase thermal control system based on a Space Station prototype thermal bus, 2) a reconfigurable computer network with alternate communications protocols, and 3) Space Shuttle Remote Manipulator System latching and deployment subsystems (7). The core ideas of CONFIG have been patented (8). CONFIG 3 has added capabilities for graph analysis and for modeling operations and procedures. Many potential uses for CONFIG in engineering and operations have been identified.

How CONFIG Can Support Engineering Activities

Major engineering activities include systems engineering (functions), physical design engineering (materials, processes, equipment performance and geometry, manufacturing, etc.), operations engineering (control, diagnostics, procedures), and supportability (reliability, maintainability) and safety (failures and hazards) engineering. There has been much progress in tool development to support systems engineering, physical design engineering, and control engineering. However, there have been missing links between physical design engineering and the other engineering areas. There has also been little software support for operations, supportability and safety engineering. Major design goals of the CONFIG tool include support for conceptual design for operations and safety engineering, and for bridging the gaps between physical design engineering and other types of engineering. These types of support will be discussed in the sections that follow in the paper.

A major area of potential CONFIG 3 use is for concurrent or integrated engineering that addresses operability, diagnosability and supportability. Operations procedures and software can be modeled along with the system model, to evaluate specific procedures, protocols, rules or plans, when they are applied in various system configurations and scenarios. Simulations of components and their interactions during operations can show how the system realizes functional requirements. Functional requirements can be embodied in operations models. Approaches to sensor location can be evaluated for support of test and diagnosis requirements. Redundancy, observability and composite measurements can be investigated in operations scenarios and fault management scenarios.

Another major area of potential CONFIG 3 use is for design definition, evaluation and documentation. Component-connection models are well suited for configuration and connectivity analysis, and for analysis of potential interactions among system parts. Graph analyses can be used to investigate completeness, consistency, modularity, efficiency, redundancy, fault tolerance and criticality. Graph analyses can also be used to locate potential failure impacts and sources. Discrete event simulation is well suited for predicting critical system states in nominal operation or when one or more failures occur. Such simulations can be used to compare alternative designs to each other and to specifications, and may include statistical studies. For design documentation, CONFIG can be used to capture functional goals in component and operations models.

CONCEPTUAL DESIGN SUPPORT

A major goal of the CONFIG project is to support conceptual design for operations and safety engineering. Major tasks in conceptual design are design definition, evaluation (by simulation and analysis) and documentation. In operations engineering, the focus is on the design of systems and procedures for operating, controlling and managing the system in normal or faulty conditions. In safety engineering, the focus is on prevention of hazardous effects and conditions in the physical system or its operation. In these types of engineering, complex interactions and interfaces among system components and operations must be a focus. This is so because of interest in such issues as operations efficiency and completeness, redundancy, fault tolerance and diagnosability. Yet, conventional physical simulation modeling provides quantitative and geometric detail where it is not needed, and leaves out modeling that

is needed of operations, faults, and component interactions. Systems engineers can model functional components and their connections, but are not supported in modeling a corresponding physical design and operations design that together achieve these functions.

Component-connection representations are well suited for modeling and defining both physical system designs (as structures of interacting components) and operations designs (as structures of interacting actions), as well as the interactions between system components and operational actions. Discrete event models have been used for this type of modeling in areas that focus on queueing and scheduling problems, but can be extended to support conceptual modeling in operations and safety engineering. This type of modeling is also compatible with systems engineering function diagrams (1).

When executable component-connection models define conceptual designs, they provide rich design documentation that can achieve most design knowledge capture goals. Design decisions can be specified by modeling operations and systems jointly, thereby showing functional intent. Design alternatives can be documented as executable models and their evaluation results. Functional intent can also be explicitly documented in operations activity models.

Discrete event simulation technology can be adapted for this purpose by incorporating abstracted modeling of component behavior and operating procedures. Such enhancements should accommodate modeling of both nominal and faulty component operating modes and both nominal and recovery operations. These modeling capabilities can be achieved by applying process modeling approaches from qualitative modeling and plan representation approaches from planning in artificial intelligence research.

The principal types of simulation and analysis enhancements needed are global capabilities, since the discrete event simulation approach is limited to local propagation of discrete change events through a component-connection structure. These global capabilities can be achieved by applying graph analysis techniques to the global component-connection structure during simulation, and in static analyses.

INTEGRATING TYPES OF ENGINEERING

Another goal of the CONFIG project is to help bridge the gaps between physical design engineering and other types of engineering, especially operations engineering and systems engineering. Such gaps impede progress in integrating engineering. The most important cause of these gaps is lack of modeling approaches that could support clean mappings among models. The CONFIG project integration philosophy is support for loose coupling among engineering support tools, so that relevant information from one tool can be selected, dispatched and mapped or translated for use by another tool. Although a single data base of interrelated models is not needed for integration, mappable modular information is needed from each tool, and distributed operation of heterogeneous tools should be assumed. Attention also needs to be paid to how selected information can be reused and how consistency can be maintained. Since many such tools, including CONFIG, use model libraries, such issues can be dealt with at both the library and the application level.

Some of these issues have been addressed in CONFIG design, to support internal integration of modeling approaches and separation of modeling and analysis concerns. Since CONFIG is intended to support incremental modeling and modeling options, modularity of models and analyses is supported. This modularity is supported by the object-oriented approach and explicit interface definitions among model types and between models and their graphical presentations.

To integrate with other types of engineering, relations between notations and representations need to be identified. Such relations can form a basis for the activities of selection, translation and mapping that support coordinated and integrated engineering. To illustrate integration with physical design engineering, we use a process engineering example (11). In process engineering, component-connection models can correspond to the process plant structures that are used in the sequential-modular approach to generating flowsheeting simulations. Likewise, components correspond to equipment, which is modeled next.

Component-connection models are closely related to functional diagram notation in systems engineering. In addition, functional information, in the form of goals for actions, is central to operations models in CONFIG. These goals describe variables that correspond to states in the system model components. Mapping to systems engineering models should not be difficult.

In operations and support engineering, tools are emerging for scheduling, planning, and representing procedures. The integration problem is to support conceptual design of operations so that the results can be reused in these tools. Operations modeling in CONFIG has been designed to correspond to current planning representations, by including goals in the activity models to achieve or maintain states. Thus, for example, CONFIG could be used both to evaluate the output of a planning or procedure development tool, and to provide a problem to a planning tool. These same capabilities could be used during operations, for evaluation of design changes for procedures, plans or schedules in the context of the current situation.

CONFIG 3

The focus of CONFIG 3 work has been on preparing and demonstrating a solid foundation for both product development and further integration studies. CONFIG has been reimplemented in a standard object-oriented language, in modular and well-documented form. The project approach has been to incrementally integrate advanced modeling and analysis technology with more conventional technology. The prototype integrates qualitative modeling, discrete event simulation and directed graph analysis technologies for use in analyzing normal and faulty behaviors of dynamic systems and their operations.

CONFIG 3 has been designed for modularity, portability and extensibility. A generic directed graph element design has been used to standardize model element designs and to promote extensibility. This directed graph framework supports integration of levels of modeling abstraction and integration of alternative types of model elements.

CONFIG provides intelligent automation to support nonprogrammer and nonspecialist use and understanding. CONFIG embeds object-oriented model libraries in an easy-to-use toolkit with interactive graphics and automatic programming.

Enhanced Discrete Event Simulation Capabilities

In traditional discrete event modeling and simulation, state changes in a system's entities, "events", occur discretely rather than continuously, and occur at nonuniform intervals of time. Throughout simulation, new events are added to an event list that contains records of events and the times they are scheduled to occur. Simulation processing jumps from one event to the next, rather than occurring at a regular time interval. Computation that results in creation of new events is localized in components, which are connected in a network. Any simulation run produces a particular history of the states of the system, and random variables are used in repeated runs to produce distributions of system output variables. These statistical simulation experiments are used to compare design alternatives.

To enhance this discrete event simulation approach to accommodate abstracted qualitative modeling of continuous behavior of system components, a number of new concepts and methods were developed for CONFIG. These concepts and methods include a component model with operating modes, types of links connecting components ("relations" and "variable clusters"), new state transition structures ("processes"), methods for representing qualitative and quantitative functions ("process language"), and a new simulation control approach.

These enhancements make discrete event simulation techniques available for evaluating conceptual designs for systems and their operations. Engineers can investigate how system components will interact in operations scenarios, in which some components can be nominal and some can be faulty, and in which effects of single or multiple faults can be local or can interact and propagate through the system. Simulations can be used to see whether system designs meet functional or redundancy requirements. Simulations can also be used to investigate alternatives for instrumenting the system, and for detecting and diagnosing faults.

Digraph Analysis Capabilities

The CONFIG Digraph Analyzer (DGA) makes graph analysis techniques available for evaluating conceptual designs of systems and their operations. The DGA is based on reachability search. This search is implemented generically so that it can be applied to any of the many types of graph data structures in CONFIG. The DGA user may specify constraints that limit the search in various ways. The results of the reachability analysis may be written to a file, presented as a textual display, or the paths found may be highlighted on an iconic screen display of the graph. In textual output mode, the DGA may also display metrics associated with the graph topology such as path lengths. Since the DGA is generic, it can be used on simplified component-connection models of systems or operations before detailed modeling has been completed.

The ability to impose constraints on the graph search allows the user to tailor analyses for a wide range of purposes. Analyses of completeness, consistency and modularity can be supported, such as ensuring that all electrically powered devices in a model are on an electrical circuit. Analysis of failure sources and impacts can be done by tracing the paths of impact of a given failure source.

System Modeling

Devices are the basic components of a CONFIG system model, which are connected together in topological model structures with Relations. Device behavior is defined in operating and failure Modes, which contain mode dependent and mode transition processes. Modes are connected together in a mode transition diagram which delineates the transition dependencies among the individual modes. Device Processes define change events in device variables, which are conditionally invoked and executed with appropriate delays during a simulation. In terms of qualitative process theory (4), a change in a component variable or a mode can be equivalent to passing a landmark value and reaching a new qualitative range. Processes define time-related behavioral effects of changes to device input variables, both direction of change and the new discrete value that will be reached, possibly after a delay. Faults and failures can be modeled in two distinctly different ways. Failure modes can be used to model device faults. Mode-transition processes can be used to model device failures that cause unintended mode changes.

Relations connect devices via their variables, so that state changes can propagate along these relations during simulations. Related variables are organized into variable clusters, to separate types of relations by domain (e.g., electrical vs. fluid connections). Relations can also connect Devices with device-controlling Activities in operations models.

Flow Path Modeling

There are two inherent difficulties in modeling flows by means of CONFIG device processes. First, processes in CONFIG are by definition local descriptions of a device's behavior while flow is in fact a global property of the modeled system and the substances being subjected to flow within the system. Second, in many cases a modeled system can undergo dynamic changes in topological structure during the course of its operations, while any process descriptions involving flow must often rely on assumptions that some aspects of the device's relationship to the system topology are static. These factors severely limit the reusability of device descriptions to a limited set of possible system topological structures.

A flow-path management module (FPMM) was implemented to address these problems, by interrupting simulation to operate on a global flow path layer of the model. The FPMM is separate from the module implementing local device behavior, but the two modules are interfaced via flow-related state variables in the modeled system's component devices. During simulation, the behavior modeling module notifies FPMM of local changes in a device's state and FPMM recomputes the global effects on flows produced by the local state change. The FPMM then updates the state of flow in all device affected by the recomputation, and this in turn may cause other processes to be invoked that result in further local changes. This interface design makes it possible for the user to write local device process descriptions that do not depend on any assumptions concerning the system topology but yet are capable of describing the mutual dependencies between the device and the system flows. Therefore, these process descriptions are highly reusable.

For large models, it would not be feasible to examine each device in a system every time any one of them underwent a process-induced state change during a simulation. FPMM therefore constructs a simplified representation of the system as a collection of aggregate objects referred to as "circuits." The devices within a given circuit are not manipulated by FPMM unless the flow state of the circuit has changed.

The complexity of the algorithm for processing the effects on system flows due to one device state change is the product of the average number of devices per circuit times the average number of circuits per device. Unfortunately, the number of circuits is itself a nonpolynomial function of the average number of connections to a device in the system (degree of the node). To increase the size of systems for which the algorithm's complexity is tractable, a second class of aggregates referred to as "clusters" was introduced. Clustering reduces a graph to a hierarchy of alternating serial and parallel clusters (6). The complexity of flow computations is a linear function of the average degree of the nodes for cluster-based representations. There are many practical examples of systems that are only partially reducible to clustering representations. For such systems, FPMM produces a hybrid representation consisting of a set of circuit objects in which clusters are treated as individual nodes. This hybrid approach can represent any arbitrary system topology, unlike pure clustering, and will allow considerably more efficient flow computations for most topologies than would pure circuit representations.

The "circuits", which are component configuration representations, simplify and separate analysis of state changes that produce changes in global configuration of flow. In many cases, configuration determination alone can be sufficient to verify flow/effort path designs, to establish flow paths for a continuous simulation, for reconfiguration planning, and for troubleshooting analysis (see Ref. [2] on cluster-based design of procedures for diagnosis, test, repair and work-around in a faulty system).

Operations modeling

Activities are the basic components of a CONFIG operations model, which are connected together in action structures with Relations. They represent procedures or protocols that interact with the system, to control and use it to achieve goals or functions. Each activity model can include evaluable specifications for what it is intended to achieve or maintain. Activity behavior is defined and controlled in a sequence of phases, ending in an evaluation of results. Activity behavior is defined by processes that simply model direct effects of actions, or that control device operation and mode transitions to achieve activity goals. Relations define sequencing and control between activities and connect Devices with device-controlling Activities.

Operations models are designed to support operation analysis with procedure models. These models are designed to support analysis of plans and procedures for nominal operation. They are also designed to support simulation and analysis of proposed design changes (reconfigured systems and revised procedures) that are developed during operations in response to failures. The procedure modeling elements are designed for reuse by intelligent replanning software, and for compatibility with functional modeling in systems engineering.

Model Development & Integration Capabilities and Approach

CONFIG provides extensive support for three separable yet tightly integrated phases of user operation during a modeling session: Library Design, Model Building, and Simulation and Analysis. This includes a graphical user interface for automated support of modeling during each of the phases including the development of object-oriented library element classes or templates, the construction of models from these library items, model inspection and verification, and running simulations and analyses.

The integration between the phases enables an incremental approach to the modeling process by allowing the user to repeatedly and rapidly incorporate into a perhaps initially simple model, lessons learned from a simulation or other analysis of that model or combination of models. This information might be used to explore a range of progressively more detailed, or simply different, structural, behavioral, and functional modifications to the various types of CONFIG models such as: addition, deletion, and reconfiguration of elements of the model layout; substitution of functionally similar elements; selective modification and redefinition of the structure and behavior of the classes from which the elements in a model were instantiated. Additionally, CONFIG's support for these phases as separate user activities fosters the achievement of concurrent engineering goals by allowing library definition, model building, and model analysis to be performed by different individuals at different times depending on area of expertise and availability of resources. Finally, support for the model building phases spans all types of modeling that can be performed in CONFIG including component structure, behavior and flow, and activity goals and structure.

The various types of model elements are instances of class definitions which are located in libraries, either CONFIG-provided or user-defined. These libraries are themselves objects which are hierarchically organized to utilize the full benefits of the object-oriented approach including inheritance. The library builder may construct one library by accessing the elements of the superlibrary and creating new elements as subclasses of the superlibrary's elements. Thus, a fundamental aspect of the automated modeling process in CONFIG is its support for creating, modifying, and storing libraries of classes, or model element templates, from which models may be constructed.

Initially the library designer may create his own object-oriented and extensible Qualitative Process Language of Variable Types, Operators, and specialized Operations which will subsequently be used to describe Device level behavior. This ability to define a domain-specific language is an important feature of CONFIG in that it does not restrict the modeler to some particular modeling language and allows him to describe the system easily using his own qualitative or quantitative vocabulary. Additionally, the library designer may then go on to create specialized subclass hierarchies of a number of other different types of CONFIG elements including Devices, Relations, Variable Clusters, and Activities.

The Device Class hierarchy illustrates how CONFIG uses the power of object-oriented definition of model elements. A Device Class defines a template from which a specific qualitative model of device behavior may be instantiated or stamped out. Device Classes not only define attributes or variables for which all instances of the class will provide instance-specific values, but also Modes of operation consisting of Mode Dependent and Mode Transition Processes that are conditionally invoked and executed with appropriate delays during a simulation. The various modes of operation are grouped together in the Mode Transition Digraph, a composite object, which may be incrementally modified as it is inherited down the Device Class hierarchy. The Processes associated with Modes contain Statements which are written by the modeler making use of the domain-specific Qualitative Process Language.

Creation of various configurations of system models may then proceed graphically and interactively. The model builder makes use of a model building window in which instances of Device and Relation Classes from a particular library are selected via mouse interaction from palettes. These instances are connected and arranged on a design canvas. Finally, CONFIG's graphical support for graph reachability and flow analysis and the running of simulations may then be used. Digraph analysis may be used on models with simpler device behavior definitions, and discrete event simulation can use more detailed device models.

Although CONFIG has been designed as a complete object-oriented qualitative modeling and simulation system which is able to stand alone with communication proceeding through use of its graphical user interface, an additional goal has been to provide the ability to integrate CONFIG with other CAE tools and data base management systems. This goal has been achieved and is reflected in the implementation of the CONFIG Programmatic Interface (CPI), which defines a set of functions and protocols for interacting with CONFIG. The CPI provides an avenue for other systems to interact with CONFIG, and for integration with object-oriented data bases.

Hosting

Care was taken to select software platforms for CONFIG that are portable to most Unix work stations. The Common Lisp Object System (CLOS) was selected as the software platform for this reason. CLOS is a highly standardized language, and two vendors produce Lisp compilers for most of the commonly available work stations. The user interface was implemented using Common Lisp Interface Manager (CLIM), another standardized tool built on CLOS and available from the same vendors. The most recent version of CLIM is designed to exploit the resources of the particular windowing system being run by the host machine so that the "look and feel" of CONFIG can be familiar.

Areas of future work

There are several areas for enhancement that we are planning to pursue. One is to provide a more complex and complete operations modeling capability. Another is to further enhance the discrete event simulation capabilities to include facilities for managing and documenting simulation experiments. We plan to integrate with an object-oriented data base management system. We also have plans for improving and enhancing flow path management and digraph analysis capabilities.

CONCLUSIONS

The CONFIG prototype demonstrates advanced integrated modeling, simulation and analysis to support integrated and coordinated engineering. CONFIG supports qualitative and symbolic modeling, for early conceptual design. System models are component structure models with operating modes, with embedded time-related behavior models. CONFIG supports failure modeling and modeling of state or configuration changes that result in dynamic changes in dependencies among components. Operations and procedure models are activity structure models that interact with system models. The models support simulation and analysis both of monitoring and diagnosis systems and of operation itself. CONFIG is designed to support evaluation of system operability, diagnosability and fault tolerance, and analysis of the development of system effects of problems over time, including faults, failures, and procedural or environmental difficulties.

CONFIG has many attributes that aid commercialization. The core CONFIG concepts are patented. CONFIG integrates advanced technology with mature discrete event simulation and digraph analysis technology bases. In the years of work on CONFIG, a number of requirements have been discovered and a number of technical and product problems have been solved. The prototype and its design are well documented, to ease conversion of all or part of the design to a supported product. CONFIG provides hooks and placeholders for further enhancements. CONFIG takes advantage of sophisticated capabilities in object-oriented databases and graphical interfaces. Commercial versions of these technologies appear to be mature enough now to support this type of advanced CAE tool.

There are several possible commercialization approaches for CONFIG. One is to simply develop a commercial version of the CONFIG tool. Another is to enhance an existing tool for object-oriented modeling or discrete-event simulation. Another approach is to integrate CONFIG with a Process simulation or Control engineering tool in a CAE environment.

ACKNOWLEDGEMENTS

The authors wish to thank Bryan Basham for significant design and coding contributions to all aspects of the CONFIG 3 prototype, prior to his leaving the project. We also thank Leslie Ambrose, Ralph Krog and Debra Schreckenghost for their contributions to user interface design, Brian Cox for his contributions to discrete event simulation design, Daniel Leifker, for his contributions to operations modeling design, and Sherry Land, for her contributions to digraph analysis design. We also thank Kathy Jurica for her continuing management support.

REFERENCES

1. Alford, M. Strengthening the System Engineering Process, Engineering Management Journal, Vol. 4, No. 1, March, 1992, pp 7-14.
2. Farley, A. M. Cluster-based Representation of Hydraulic Systems, Proc. 4th Conference on AI Applications, March, 1988, pp. 358-364.
3. Fishman, G. S. Principles of Discrete Event Simulation. New York, NY: Wiley, 1978.
4. Forbus, K. Qualitative Physics: Past, Present, and Future. In Exploring Artificial Intelligence (H. Shrobe and AAAI, eds.). San Mateo, CA: Morgan Kaufmann, 1988.
5. Fullwood, R. R. and Hall, R. E. Probabilistic Risk Assessment in the Nuclear Power Industry: Fundamentals and Applications. Pergamon Press, 1988.
6. Liu, Z. and Farley, A. M. "Structural Aggregation in Common-Sense Reasoning". Proc. 9th National Conference on Artificial Intelligence (AAAI-91), July, 1991, pp. 868-873.
7. Malin, J. T., B. D. Basham and R. A. Harris, "Use of Qualitative Models in Discrete Event Simulation for Analysis of Malfunctions in Continuous Processing Systems." Artificial Intelligence in Process Engineering (M. Mavrovouniotis, ed.), Academic Press, pp. 37-79, 1990.
8. Malin et al., U. S. Patent 4,965,743, "Discrete Event Simulation Tool for Analysis of Qualitative Models of Continuous Processing Systems" October, 1990.
9. Malin, J. T., and Lance, N. "Processes in construction of failure management expert systems from device design information". IEEE Trans. on Systems, Man, and Cybernetics, 1987, SMC-17, 956-967.
10. Malin, J. T. and Leifker, D. B. "Functional Modeling with Goal-Oriented Activities for Analysis of Effects of Failures on Functions and Operations". Informatics and Telematics, 1991, 8(4), pp 353-364.
11. Winter, P. Computer-Aided Process Engineering: The Evolution Continues. Chemical Engineering Progress, February, 1992, pp 76-83.

omit

COMPUTER HARDWARE

SPACECRAFT ON-BOARD INFORMATION EXTRACTION COMPUTER (SOBIEC) P 10

David Eisenman - Deputy Section
Manager, Flight Command &
Data Management Section
NASA Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109

Robert E. DeCaro
Chief Electronic Systems Engineer
Irvine Sensors Corporation
3001 Redhill Avenue
Costa Mesa, CA 92626

David W. Jurasek
Director of Advanced Systems
Engineering
nCUBE
1825 N.W. 167th Place
Beaverton, Oregon 97006

ABSTRACT

The Jet Propulsion Laboratory is the Technical Monitor on an SBIR Program issued for Irvine Sensors Corporation to develop a highly compact, dual use massively parallel processing node known as SOBIEC. SOBIEC couples 3D memory stacking technology with state of the art parallel processor technology provided by nCUBE. The node contains sufficient network Input/Output to implement up to an order-13 binary hyper-cube. The benefit of this network, is that it scales linearly as more processors are added, and it is a superset of other commonly used interconnect topologies such as: meshes, rings, toroids, and trees. In this manner, a distributed processing network can be easily devised and supported. The SOBIEC node has sufficient memory for most multi-computer applications, and also supports external memory expansion and DMA interfaces. The SOBIEC node is supported by a mature set of software development tools from nCUBE. The nCUBE operating system (OS) provides configuration and operational support for up to 8000 SOBIEC processors in an order-13 binary hypercube or any subset or partition(s) thereof. The OS is UNIX (USL SVR4) compatible, with C, C++, and FORTRAN compilers readily available. A stand-alone development system is also available to support SOBIEC test and integration.

MISSION REQUIREMENTS

The general problem of finding optimal techniques for the extraction of scientific information from a wide band data stream has been discussed in depth in a JPL publication by Robert Rice¹. There, the observation was made, and to a degree quantified, that perhaps the most powerful technique for error-free information extraction is to employ activity and pattern recognition to cue the allocation of digitization and communications resources. In discussions with JPL personnel regarding this technique, the example was given of a Mars explorer spacecraft

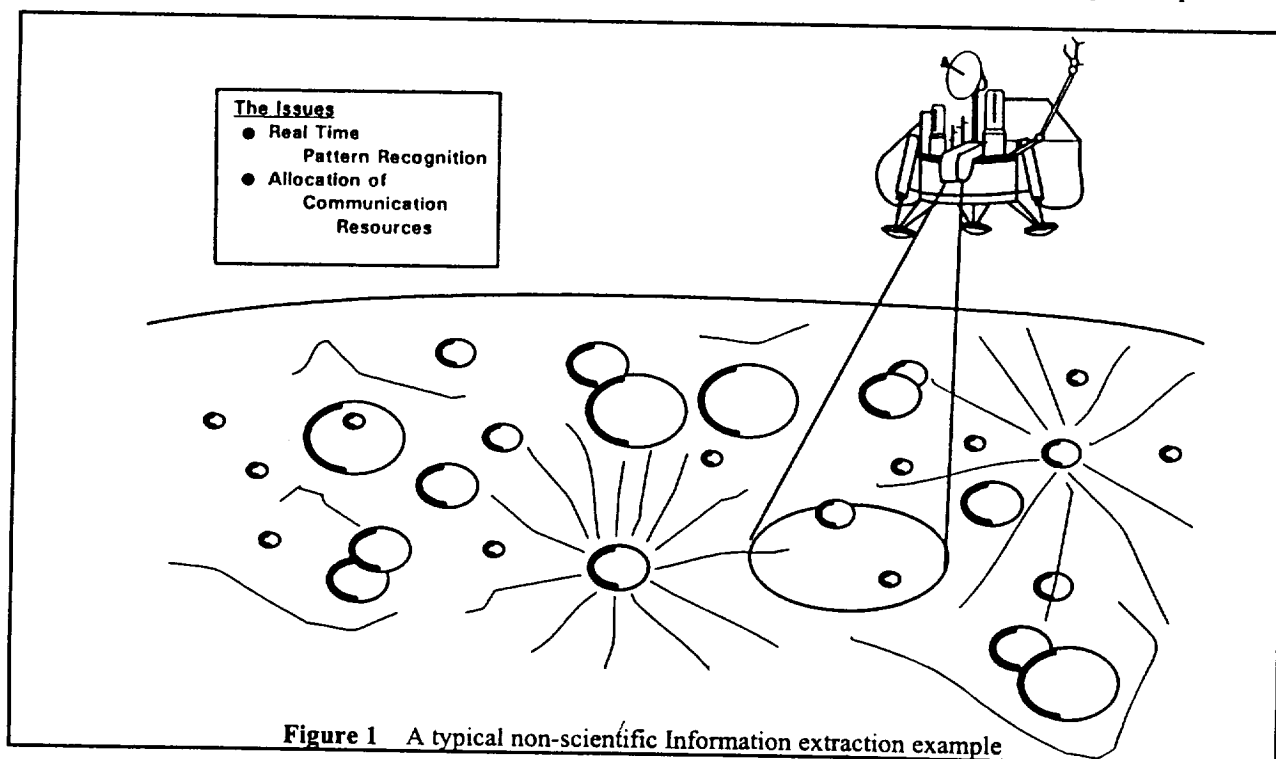


Figure 1 A typical non-scientific Information extraction example

looking for a landing site (Figure 1). In the Mars lander case, it is desirable to avoid areas of high activity and spatial complexity. It is necessary to examine apparent clear areas in very great detail (high spatial and high amplitude resolution) to assure that these areas are indeed clear and flat. In this example, sophisticated feature extraction and pattern recognition capability is important. Comparison of this example to the more obvious one of high fidelity scientific data communication in the face of a limited datalink provides evidence of the generality of the information extraction problem. A general solution to this problem is the Spacecraft on-Board Information Extraction Computer (SOBIEC), a massively parallel, highly interconnected processing system. This effort is funded by a Small Business Innovation & Research (SBIR) contract, monitored by the Jet Propulsion Laboratory (JPL).

Figure 2 shows a concept diagram for using a high-density, parallel processing computer with a large amount of distributed memory to perform feature extraction, which leads to a prioritized downlink of important features at high resolution and optimizes the limited bandwidth communication channel.

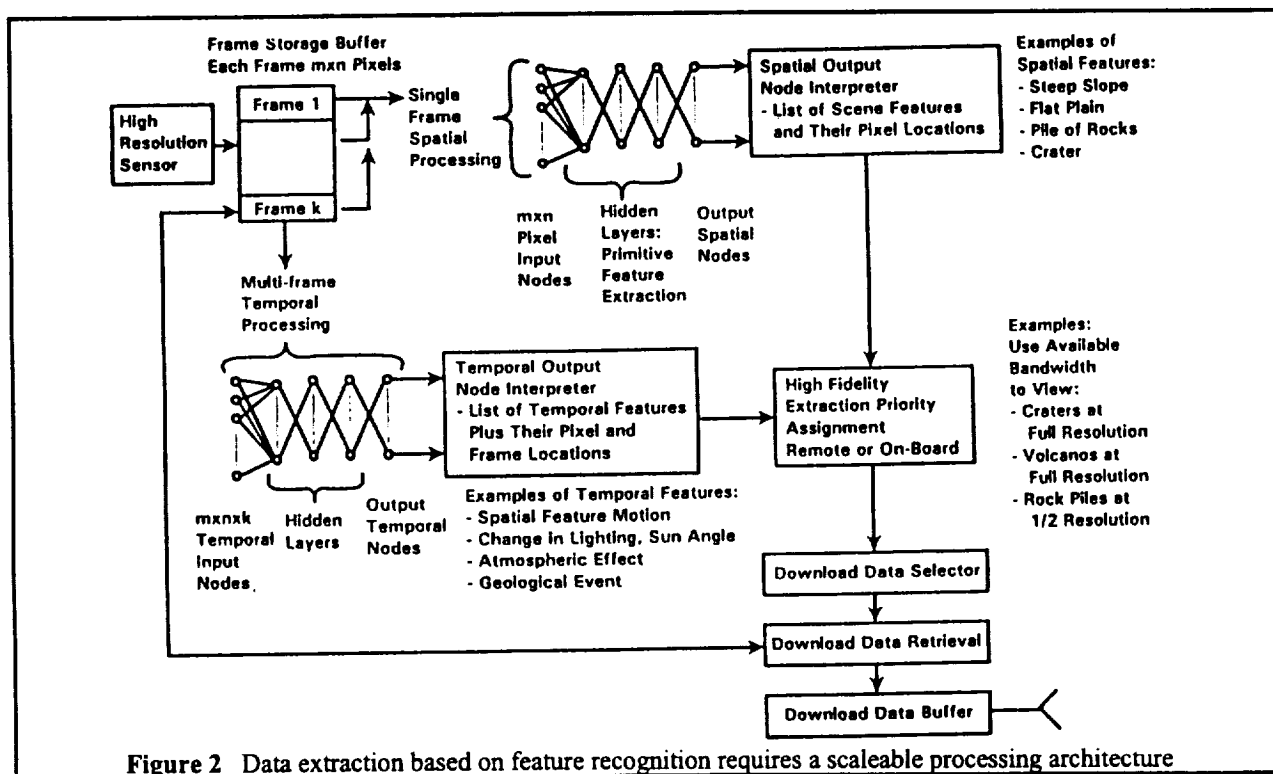


Figure 2 Data extraction based on feature recognition requires a scalable processing architecture

This implementation of feature extraction uses parallel processors to emulate neural circuitry performing hierarchical pattern recognition. For any given mission, both hardware utilization (number of nodes and interconnections between nodes) and software (weight definition specific to features require definition. These definitions are left open to make the implementation generic. Nevertheless, it illustrates a potentially powerful method for feature recognition and image extraction, which is also well suited for hardware implementation on a distributed memory parallel processing computer. The SOBIEC massively parallel processing node, developed concurrently between Irvine Sensors Corporation (ISC), nCUBE, and NASA JPL, a highly compact building block, enables compute intensive missions where the processing must be scaled to the application such as the example given, micro-spacecraft and micro-rovers.

SOBIEC ARCHITECTURE

SOBIEC's electrical architecture, typical of massively parallel processors, is shown in Figure 3. SOBIEC's processor (developed by nCUBE) contains a dynamic RAM controller with 7 bit error detection and correction (EDAC) and fourteen serial communications links. Ten years ago, nCUBE pioneered the field of massively-parallel computing where hundreds or thousands of processors are used to solve large, complex computing problems.

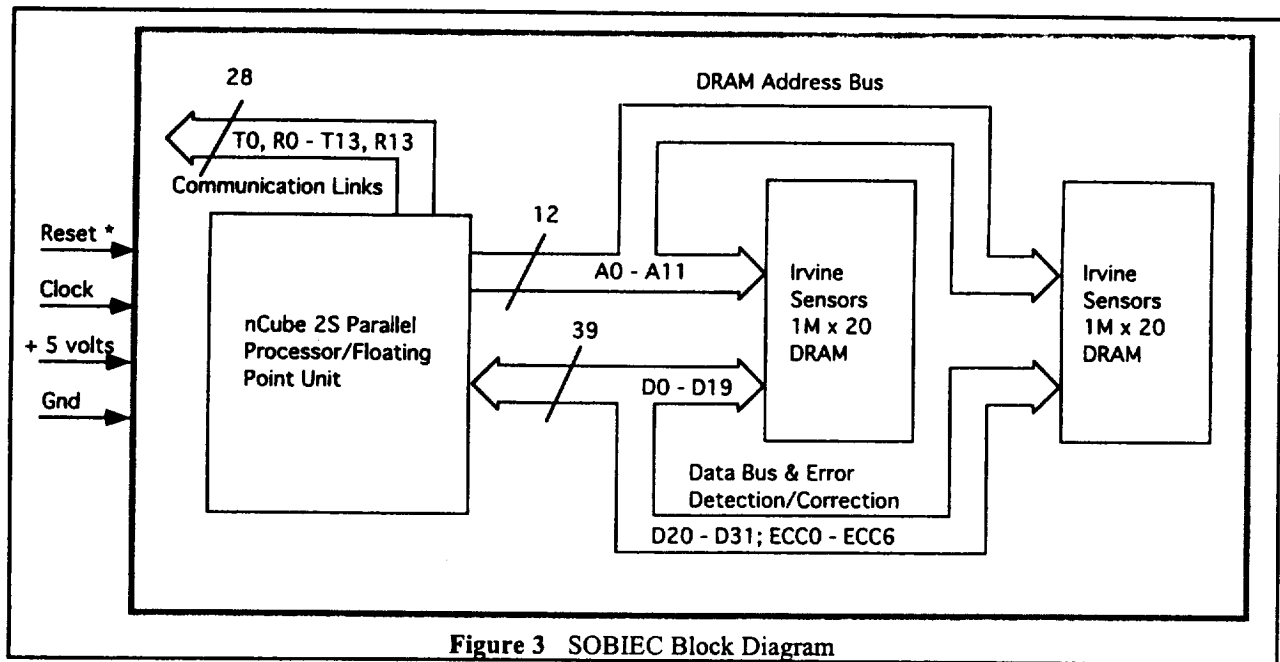


Figure 3 SOBIEC Block Diagram

An important aspect of parallel computers is their ability to continue operating, even when one or more processing elements has failed. Although the failure rate of SOBIEC's highly-integrated processor is extremely low, this "graceful degradation" in computing is vital to mission-critical applications. The relatively low cost of the SOBIEC processor node allows redundancy to be used in a cost-effective manner when needed, while still allowing the flexibility to re-allocate resources to handle peak loads.

SIMD vs. MIMD PARALLEL COMPUTERS

Over the years, many different parallel computer architectures have been proposed and built. Generally, these machines fall into two main categories: Single Instruction, Multiple Data (SIMD) and Multiple Instruction, Multiple Data (MIMD).

SIMD machines are characterized by having one set of instructions (the program) and multiple data sets. In these types of machines, the data is spread over the array of processors, and then all processors execute the same instructions in lockstep. Although this is extremely useful for certain types of applications where the same manipulation of data is applied across a large data set, it has a number of disadvantages for a SOBIEC application. First, it is virtually impossible to have multiple tasks (different applications) on these machines. Secondly, programs on SIMD machines can be difficult to debug because it is difficult to view the data on selected processors as the program is running. Finally, these machines lack flexibility. SOBIEC programs could not "interact" with the data on a processor-by-processor basis, modifying the execution of the program based on the data in an individual processor. These disadvantages proved to be a serious limitation for SOBIEC's mission, and so MIMD processors were investigated.

MIMD machines are characterized by each processor having its own individual instruction set (program) and data set. This distribution of both instructions and data gives these machines a large amount of flexibility. It is very straightforward to divide these machines among multiple tasks, each task taking only as much of the system compute power as is necessary to complete the task in a time-effective manner. Debugging on these machines is aided greatly by the fact that a debug program can be loaded on one or more selected processors, giving the programmer the ability to view the state and data on individual processors or set of processors in the machine. Finally, since separate programs are running on each processor node, these programs need only perform the operations appropriate to the data stored locally. This "extra" processing power can then be effectively used to perform other "background" tasks, increasing the overall flexibility and cost-effectiveness of the machine.

Another distinguishing characteristic of SOBIEC's processing computer is that it is based upon a distributed vs. shared memory architecture. In the shared memory system, all processors have access to common

memory pool, in which instructions and data are stored. Although the shared memory model has the advantage of being familiar to virtually all programmers, it suffers from the serious flaw of being difficult to scale. The limited data bandwidth of the busses that connect all processors to common memory quickly becomes the bottleneck of the system, preventing additional processors from providing the expected increase (linear scaling) in performance. Distributed memory systems (such as SOBIEC) give each processor its own local memory, which gets shared with other processors and the outside world via messages over a communications network. Since these are "private memory arrays", the total memory bandwidth increases linearly as more processors are added. Thus, memory bandwidth is not a limiting factor in the scalability for SOBIEC systems.

SOBIEC's (nCUBE's) communications network is known as a binary hyper-cube. In a binary hyper-cube system, all processors are assigned a binary identification word (the Processor ID). Processors which differ by only one bit in their ID are interconnected with a synchronous, duplex Direct Memory Access (DMA) channel which runs at 2.75 megabytes per second, each direction. Thus, the number of DMA channels on any given processor is the log (base 2) of the maximum number of processors in the system. SOBIEC's processor has 13 DMA ports for array interconnect, allowing up to 8192 processors to be fully interconnected. An additional (fourteenth) DMA port on each processor is used to connect to the outside world via the I/O subsystem, which would consist of additional SOBIEC processors running I/O driver code. The benefit of this network is that it scales linearly as more processors are added, and it is a superset of other commonly-used interconnect topologies such as: meshes, rings, toroids, and trees. This feature makes the SOBIEC processor node truly universal for NASA applications.

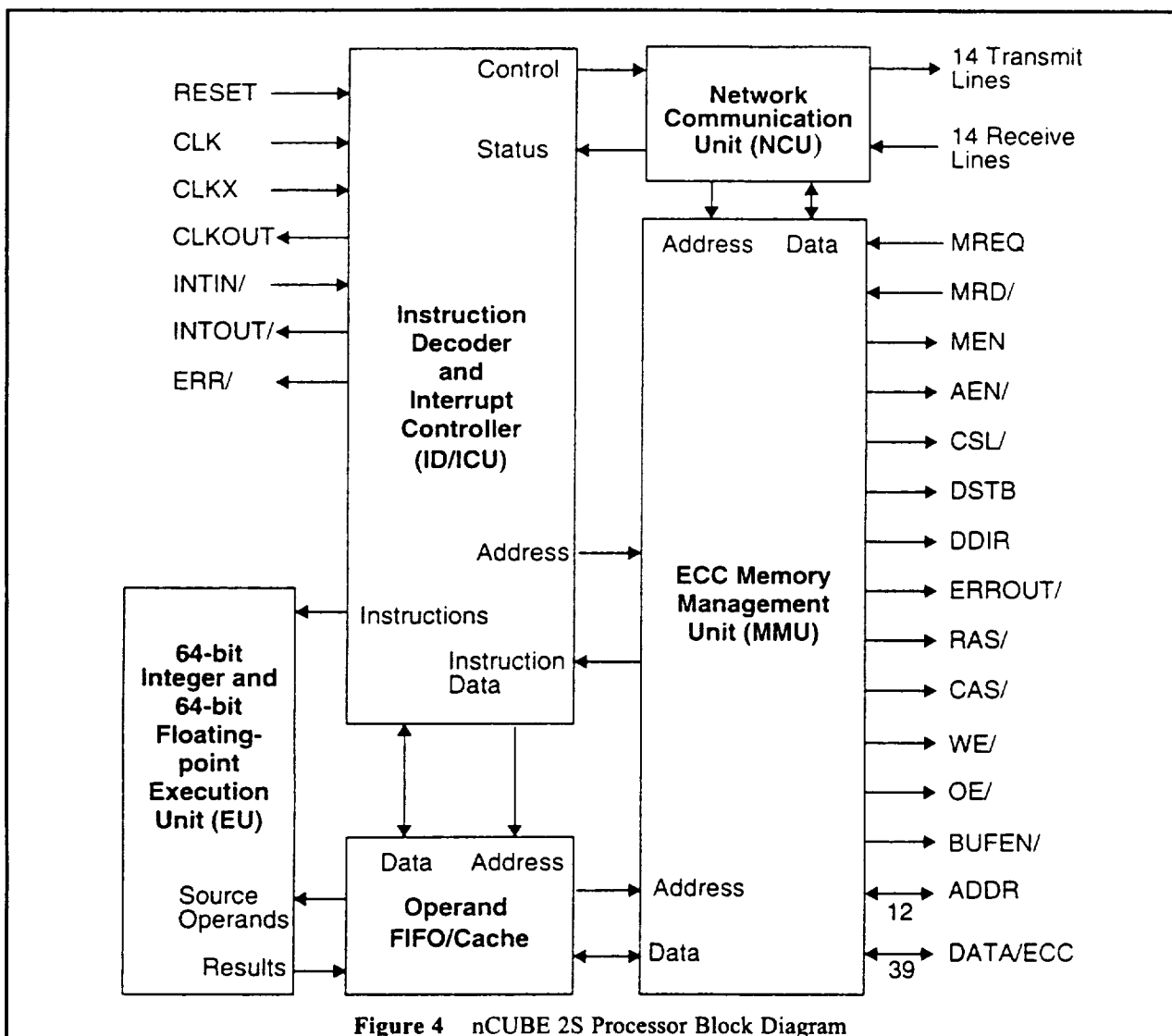


Figure 4 nCUBE 2S Processor Block Diagram

PROCESSOR SELECTION TRADES

The issue as to whether to use custom, proprietary processor chips, or off-the-shelf commercial processor chips for the SOBIEC program was explored. The advantages of off-the-shelf processors are fairly obvious. These parts are inexpensive, reliable, readily-available, and typically have wide familiarity in the marketplace. But we also found that they also have one large, fairly non-obvious disadvantage: in order to be commercially successful, the manufacturer has had to make these parts quite general purpose and able to fit into a wide variety of systems. This means that for any given system, a fair amount of "glue logic" to make the processor work is required. All that "glue" would make SOBIEC physically larger, slower, less reliable, more power hungry, etc.

nCUBE, in designing the processor used for SOBIEC, chose a more difficult path by designing its own full-custom processor. By designing a minimum parts count "world-class processor", all the disadvantages associated with commercial processors were avoided. The nCUBE MIMD processing element consists of the processor chip and DRAM (Dynamic RAM) chips... nothing else. These elements are physically small, reliable, low power and easy to build and test, making them an ideal choice for SOBIEC. Our experience has also shown that the extra money spent on a relatively low volume custom processor is more than made up for by cost savings due to the lack of glue logic, and simplified build/test of MCMs.

SOBIEC PROCESSOR FEATURES

Figure 4 is a block diagram of nCUBE's 2S processor used on SOBIEC. The 2S processor consists of five major blocks: Instruction Decoder/Interrupt Controller (ID/ICU), 64-bit Integer/Floating Point Execution Unit (EU), Operand First-In, First-Out (FIFO) Cache, Network Communication Unit (NCU), and ECC Memory Management Unit (MMU). This processor features a 64-bit data path for on-chip integer, floating point processors and dynamic Memory Management unit. The nCUBE 2S Processor is capable of addressing from 1 to 64 Megabytes of memory per node. In addition, the processor supports page mode accesses to increase bandwidth. The MMU directly controls the external DRAM chips with an external 32-bit data bus and 7-bit ECC (Error Check and Correct) bus. This ECC provides SOBIEC the assurance that all single-bit errors will be detected and corrected on-the-fly, while double bit errors will be detected and flagged as such. The extra security provided by this ECC ensures that SOBIEC's space applications will provide correct results, despite the occurrence of such well documented random events as alpha-particle hits.

The processor contains protection logic to protect system software and allow multiple processes per node. A User/Supervisor bit in the Program Status Word (PSW) causes protection faults on certain instruction and restricts access to memory other than the memory assigned to a process.

nCUBE's 2S processor contains an elaborate interprocessor communications network. Fourteen DMA channels, tightly coupled to the MMU enable SOBIEC to share the mission signal processing between on-board processing elements. An additional benefit of this communication network, is in developing a fault tolerant architecture. As stated previously, in the event of a device failure, a properly designed multiprocessor system can be designed to re-allocate resources to achieve a graceful degradation - key to any successful space-based application.

The 2S processor's high integration also results in significant power savings - about 2.5 watts at 20 MHz. Combined with four Megabytes of memory, SOBIEC's total power dissipation is only 6 watts per node. In addition, each node is capable of 3.2 MFLOPs (Million Floating-Point Operations per second), 12 MIPs (Million Instructions per second), and 80 megabytes per second of memory bandwidth.

3D STACKED MEMORIES

The SOBIEC MPP node achieves its small size by utilizing a pair of 20 megabit DRAM "short stacks" configured as a 1 megabyte x 20 bit word manufactured by Irvine Sensors, Costa Mesa, California. The memory devices are referred to as 3D silicon "short stacks" because the individual memory die are layered on top of each other, similar to a "stack of pancakes". This configuration enables a minimum height component with only a slight overall increase in height (0.060 inches versus 0.025 inches) of the original silicon. The process used to fabricate Irvine Sensors' 3D silicon "short stacks", shown in figure 5 will now be described.

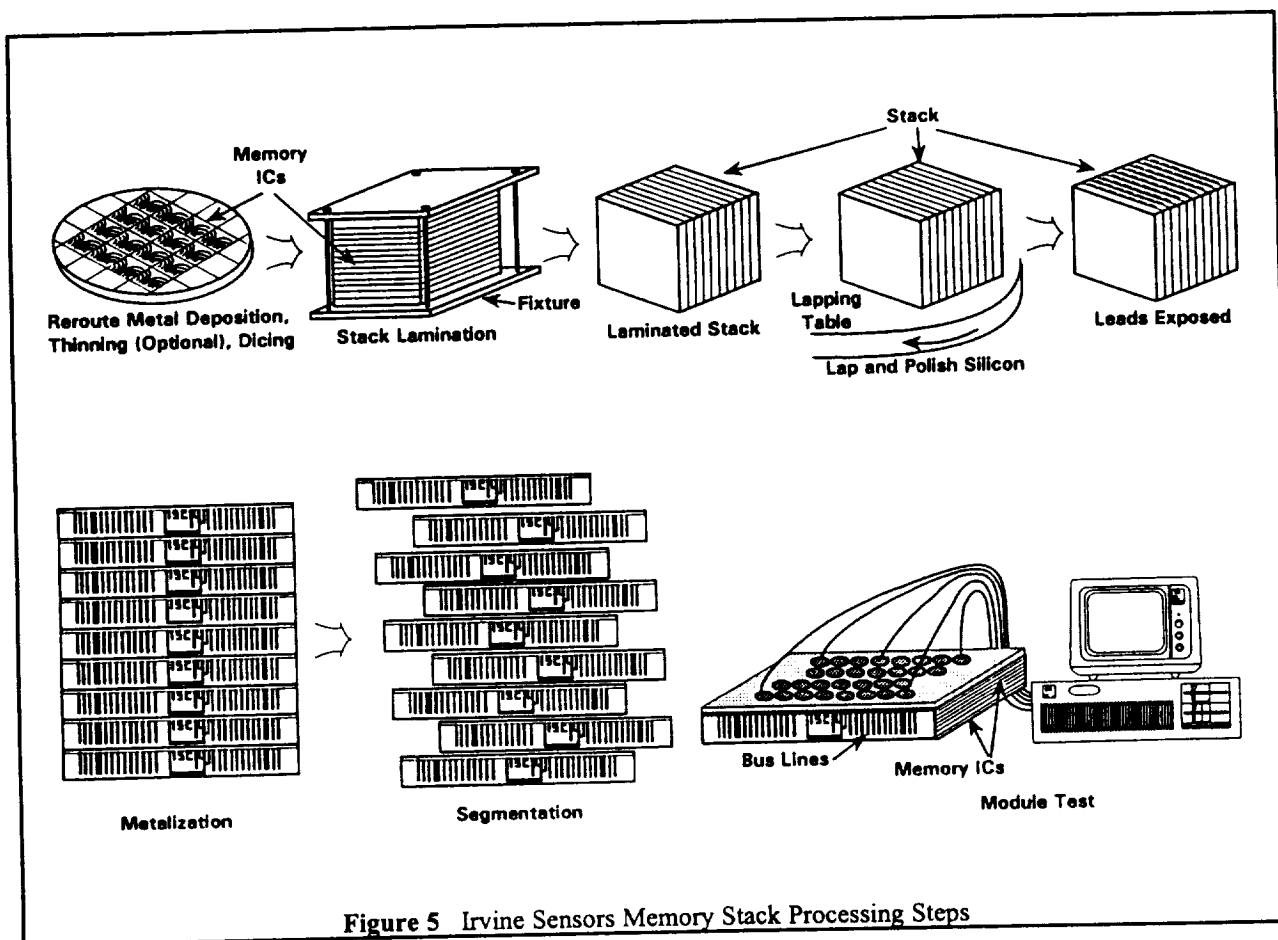


Figure 5 Irvine Sensors Memory Stack Processing Steps

Process Description

3D silicon memory fabrication begins by performing a device lead re-route at the wafer level to bring all the input-output to one side of the die. The wafers are thinned to about 0.010 inches and then sawn to provide individual memory die. The good die are laminated together (to form a cube) using a judicious application of a pair of dielectric adhesives (thermal setting and thermal plastic), special separation ceramic cap chips (top and bottom for eventual separation into short stacks) and lamination fixtures. Following cube lamination, the "cube" is lapped and polished (on the re-routed lead side) to expose the lead conductors. An etching process is applied to the cube for further lead exposure prior to passivation. The cube gets further lapping to again expose the re-routed leads (from the passivation) and then the cube gets a final (bussing) metallization applied to electrically interconnect the devices. To obtain the five layer long-word memories, the cube is heated to its thermal plastic region, and the devices are segmented forming "short stacks". The excess adhesive is then removed from the top surface, and the memory component undergoes final testing.

Technology Maturity

Irvine Sensors has spent over ten years in the development of the stacked memory technology. During that time, 128-layer stacks of 4-mil layer thickness were successfully fabricated and tested to cryogenic temperatures for infrared focal plane applications. Layers as thin as two mils were successfully fabricated. "Short stacks" have survived over 100 cycles of -55°C to +125°C long thermal cycles with no change in interlayer wiring resistance. In addition, 50-layer memory stacks of 15-mil layers have been successfully fabricated, tested and temperature cycled to over 300°C with no change in interlayer wiring resistance. The stability of the technology is so promising that IBM and Irvine Sensors entered into a joint development venture (August '92) to start a memory cubing production line in Burlington, Vermont. This line is projected to be in production of 3D memory products in 1Q94.

Advantages of Irvine Sensors 3D Memory Short Stacks

Laminated 3D memory, which is processed using thin film technology for interlayer connections, is inherently simple and robust. This technology provides maximum design flexibility and density. The advantages of Irvine Sensors 3D stacked memory are summarized in Table 1.

Table 1 Advantages of Irvine Sensor's 3D Short Stack Memory Technology

Attribute	Advantage
Height	<ul style="list-style-type: none">• Stack height is minimized by thinning ICs to a 10 mil thickness• Height with two cap chips is 60 mils (four layer memory stack)
Reliability	<ul style="list-style-type: none">• Interconnect with standard semiconductor thin film high reliability techniques• External shock & vibration do not effect interconnections due to mass & rigidity of stack
Cost & Availability	<ul style="list-style-type: none">• IBM & Irvine Sensors recently opened a high rate production facility for stacked memories• First Product Available 1Q94
Power	<ul style="list-style-type: none">• Capacitance, inductance & RFI susceptibility are reduced with 3D silicon
Speed	<ul style="list-style-type: none">• Speed is optimized due to the reduced capacitance interconnect wiring• No more direct or lower impedance interconnects possible than Irvine Sensor's stacked memory technology

EARLY SOBIEC PACKAGING CONCEPTS

The design goals for the SOBIEC program are to develop a high performance, minimum parts count distributed memory massively parallel processing node, that is: low in power, weight, and cost; small sized; and contain sufficient memory for most multi-computer applications. Further system level requirements levied were: low thermal impedance (4°C/Watt Junction to Case) to enable applications with severe temperature extremes, and low conducted noise (EMC). Our early SOBIEC conceptual approach (figure 6) was a 3D silicon architecture that utilized nCUBE's n2S single chip processor as an active substrate for Irvine Sensors' 3D silicon memory devices.

In this approach, a third layer of metallization is added to re-route the processor's DRAM compatible Input/Output to a single stack of ten 4 Mbit DRAMs configured as a 1 Megabyte by 40 bit word. The mechanical interface between the processor and memory, consisted of thermal epoxy and direct wire bonding between the memory stack and processor. Also, this approach required one dimension of the mechanical interface between the processor and memory, to be similar in length to the longest dimension of the stacked memory, in order to provide a stable base for memory attachment. After undergoing a design rule shrink however, nCUBE's processor failed to meet this requirement, and so an alternate approach was sought.

FINAL SOBIEC PACKAGE DESIGN

After several "team" meetings, a pseudo dual cavity 138 pin grid array, alumina package approach (shown in figure 7) was selected for the SOBIEC baseline design. The salient features of this approach are excellent MCM testability, low thermal impedance -- no localized processor heating, no external "glue or ancillary parts," 4 megabytes main memory upgrade able to 16 megabytes (32 megabytes possible with an increase in height of just 0.050 inches!) in the same footprint, small sized -- only 1.2 by 1.2 by 0.31 inches in height, and low electrical noise generation.

The benefits of this approach are: simple implementation of X-Y tiled arrays of processing nodes. Each "node" requires only about one third of the original space. The reduced area is a direct result of Irvine Sensors' 3D silicon memory technology. These memories are located directly under the nCUBE n2S processor, separated by 0.09 inches of alumina. Since these memories are only 0.060 inches thick, and require little more ceramic real estate than the processor itself, a highly compact massively parallel processor node was enabled.

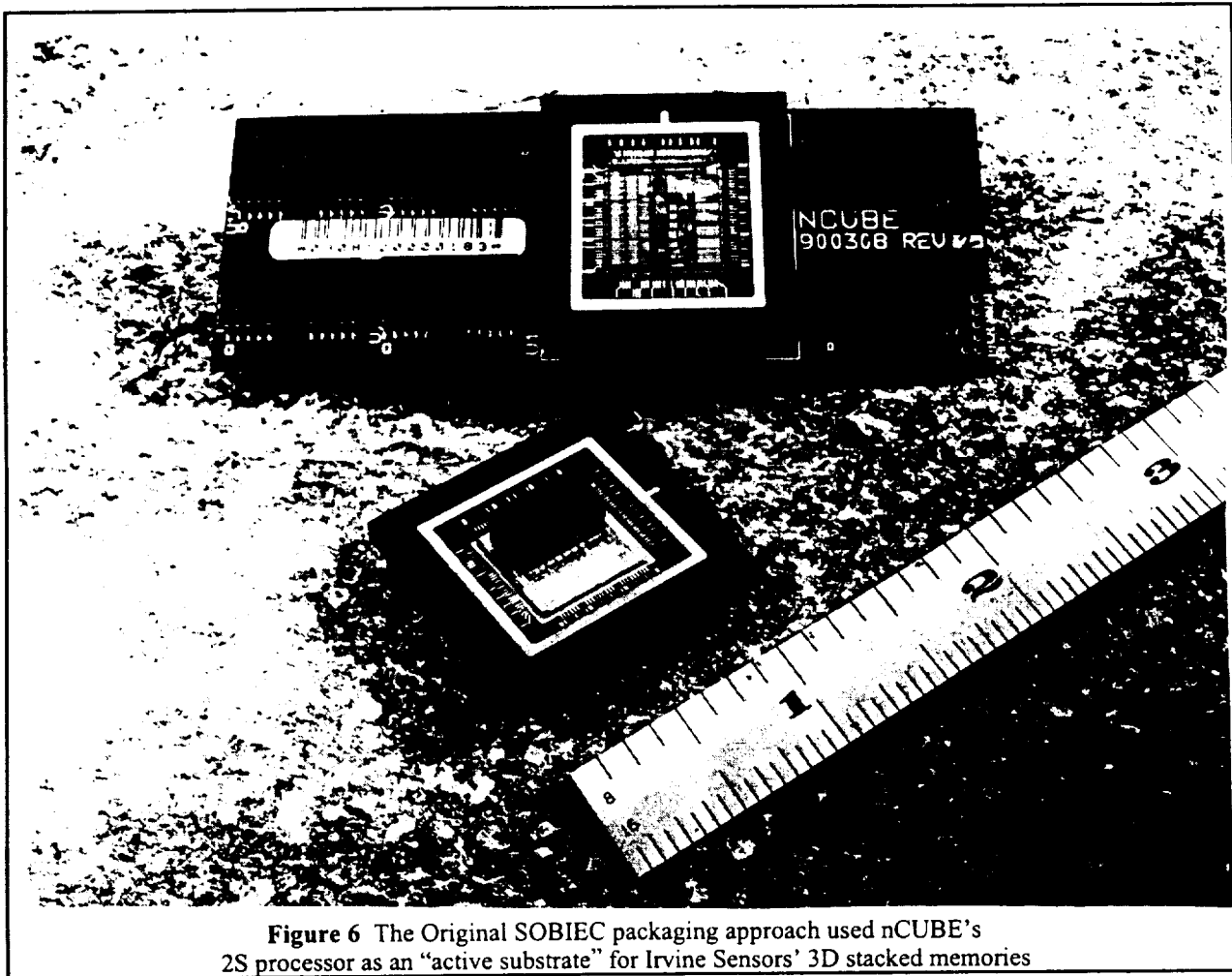


Figure 6 The Original SOBIEC packaging approach used nCUBE's 2S processor as an "active substrate" for Irvine Sensors' 3D stacked memories

nCUBE's n2S processor is implemented in 1.0 μm CMOS technology, and has modest power requirements. However its high clock frequency and numerous output buffers (address/data, control, error detection and correction) can cause "power surges" as multiple output buffers drive new signal levels simultaneously. Clean on-chip power distribution at (almost) all frequencies is provided by generous "package integral" and "package exterior" voltage bypass capacitors. The package exterior capacitors are a mixture of SMT tantalum and ceramic devices. These devices have been carefully chosen to provide less than 0.1 ohm ESR (equivalent series resistance) from about 2 kilohertz to over several tens of megahertz. Beyond this frequency, SOBIEC's "package integral" capacitance (about 0.001 microfarad parallel plate capacitor) formed by its multiple internal power and ground planes, provides effective noise bypassing to several hundred megahertz. In addition, SOBIEC input power is supplied by 38 V_{DD} and V_{SS} pins to assist in the supply of low noise power.

SOBIEC RELIABILITY

During the contract period, Irvine Sensors and nCUBE evaluated the reliability of the SOBIEC module in a 55° C environment. The reliability of the SOBIEC processor node included data analysis of the SOBIEC thermal management system using a combination of previous data and SOBIEC package thermal characteristics. The following thermal impedances were used for the evaluation:

Package Thermal Impedance	3°C/Watt	Junction To Case
Short Stack Thermal Impedance	3°C/Watt	Junction To Case
PC Board Thermal Management	10°C	Overall Case To Ambient Temperature Rise

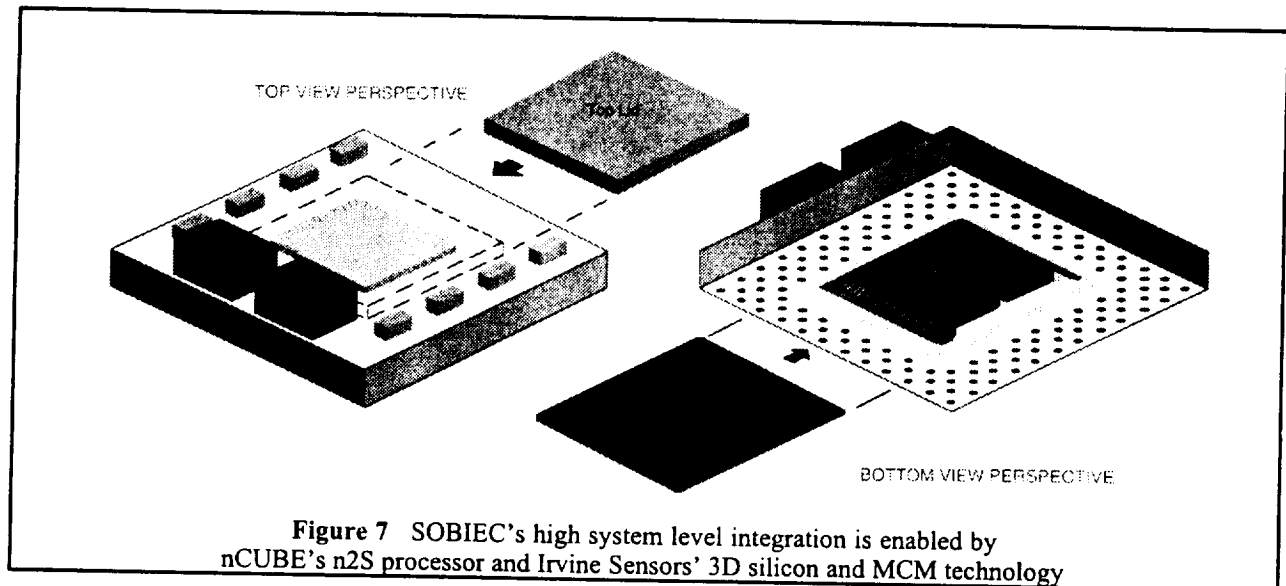
The power dissipations for the SOBIEC electronics are as follows:

Short Stack Memories (0.33 watts x 5)	1.65 watts (each stack)
nCUBE Processor	2.5 watts
Total SOBIEC Power	~ 6 watts

Using this data, SOBIEC's alumina package would elevate in temperature 18°C ($3^{\circ}\text{C/W} \times 6 \text{ W}$). In addition, the delta temperature of the top-most DRAM die in the short stack would be 5°C ($1.65 \text{ watts} \times 3^{\circ}\text{C/watt}$). Therefore the junction temperatures of the SOBIEC devices are as follows (for a 55°C environment):

nCube Processor	83°C
Short Stack Memories	88°C

This data translates to an MTBF for the SOBIEC module of about 3 million hours.



SOBIEC OPERATING SYSTEM (OS) AND SOFTWARE SUPPORT

SOBIEC acceptance by NASA is not only due to its unique packaging, but due to nCUBE's mature development toolset. nCUBE's software--the Parallel Software Environment--provides a familiar UNIX interface with extensions and optimizations to take advantage of SOBIEC's massively parallel hardware. The software includes a micro kernel, libraries, and UNIX utilities for SOBIEC (nCUBE) processors, as well as development and system management software for workstations on the nCUBE's network. Each software component in the Parallel Software Environment has been designed for speed and flexibility--the goals of massively parallel computing.

Running on every nCUBE processor, the nCX operating system manages processes, memory, and interprocessor communication, and supports a UNIX system call interface and POSIX signals -- all in a compact, optimized micro-kernel. Because nCX runs on I/O processors as well as compute processors, programmers can develop custom device drivers in the standard nCUBE programming environment.

nCUBE libraries include standard UNIX libraries, parallelization libraries, math libraries, and graphics libraries. In many cases, a programmer can port an application to an nCUBE 2 supercomputer simply by inserting a few parallelization calls. These calls hide the underlying communication necessary to parallelize an operation, and perform complex operations blazingly fast. A Fast-Fourier Transform or a matrix multiply can be admirably performed with a single call. All the libraries support striped files for faster I/O. Users can run UNIX utilities such as cat and tar on nCUBE processors. These utilities operate on striped files transparently.

The Parallel Software Environment's workstation software includes a set of cross-development tools for writing, compiling, launching, profiling, and debugging parallel programs. The tools use the interfaces and command-line options of tried-and-true UNIX tools. Using the debugging and profiling tools, programmers can step through parallel programs or generate bar graphs of subroutine usage or communication loads. The tools make it possible for UNIX programmers to quickly learn the basics of developing, debugging, and tuning parallel programs. Workstation software also includes user and system administration utilities for monitoring and managing the nCUBE 2 supercomputer. Users can control SOBIEC processes, load multiple programs in complex configurations, or display sample code for a subroutine. System administrators can track system usage with an accounting system, selecting shut down and reboot I/O servers, and manage resources with nQS, nCUBE's batch queuing system.

nCUBE is continuing to develop its software, making parallel processing and parallel I/O faster than ever. Within the year, nCUBE will introduce a new parallel file system that supersedes RAID 5 in performance and reliability. nCUBE is also continuing to develop its networking and database capabilities.

COMMERCIALIZATION POTENTIAL

The commercialization potential for SOBIEC is enormous. Recently, the commercial market has begun to benefit from the power of massively parallel computers. Parallel computing is taking a dual path to success. The first set of commercial users are strictly interested in the number of processing nodes that can be placed onto a fixed board. In this case, SOBIEC clearly has an edge of it's competition due to it's unique packaging technology. In the second case, commercial users are most interested in matching input/output bandwidth through the use of a parallel configuration of computers. Here again, SOBIEC's small physical size and high degree of interprocessor communications provides a competitive edge over similar technologies.

A commercial application that can immediately benefit from SOBIEC, involves applications requiring large databases such as Oracle. SOBIEC's 2S processor is designed to rapidly process transactions and very complex queries using Oracle. The natural parallelism of information in commercial databases makes them an ideal fit for SOBIEC's massively parallel computing.

CONCLUSIONS

A highly compact high performance massively parallel processing system has been developed by Irvine Sensors, nCUBE, and NASA JPL, and is in the final stages of integration and test. This production ready design realized significant size, weight, and volume reductions through the judicious application of 2D and 3D silicon technology. This general purpose processing element is packaged in a 138 leaded pin grid package that requires no more board real estate than the original packaged processor itself. The low cost alumina package exhibits excellent thermal and electrical properties and meets all the requirements for a SOBIEC mission. Completing the introduction of this product, is a mature software development system and library to ease the new or experienced user into the work of massively parallel computing.

References:

- 1 JPL Publication 82-61, End-to-End Imaging Information Rate Advantages of Various Alternative Communications Systems by Robert F. Rice, 9-1-82

PEN-BASED COMPUTERS: COMPUTERS WITHOUT KEYS

Cheryl L. Conklin
Systems Design Engineer
Analex Space Systems, Inc.
P.O. Box 21206
Kennedy Space Center, FL. 32815-0206

2711
P-7

ABSTRACT

The National Space Transportation System (NSTS) is comprised of many diverse and highly complex systems incorporating the latest technologies. Data collection associated with ground processing of the various Space Shuttle system elements is extremely challenging due to the many separate processing locations where data is generated. This presents a significant problem when the timely collection, transfer, collation, and storage of data is required. This paper describes how new technology, referred to as Pen-Based computers, is being used to transform the data collection process at Kennedy Space Center (KSC). Pen-Based computers have streamlined procedures, increased data accuracy, and now provide more complete information than previous methods. The end result is the elimination of Shuttle processing delays associated with data deficiencies.

INTRODUCTION

As paperwork associated with Shuttle processing continues to grow in volume, along with it grows the need to increase manpower and equipment (Figure 1). Today's budget cuts and continued reduction in the workforce made the problem seem hopeless. An innovative way to collect data in an accurate and timely manner, and at the same time provide automated validation, was needed. The Pen-Based computer has provided the solution to this problem. The system not only collects data but also provides a historical database for analysis. Through networking capabilities, this data can be made available to users at all National Aeronautical Space Administration (NASA) centers and contractor locations.

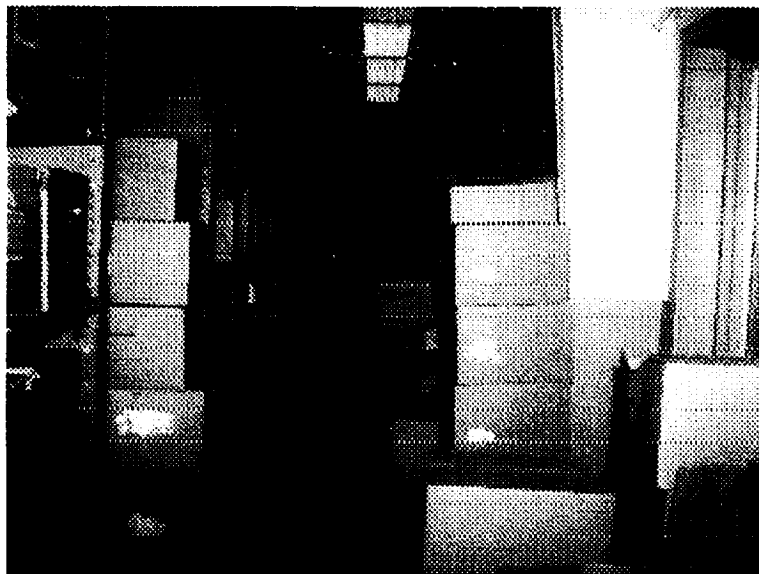


Figure 1 Data Storage Area at KSC

BACKGROUND

The existing process of collecting data at KSC is time consuming, labor intensive, and cannot provide real-time support. These deficiencies become extremely critical when dealing with shuttle quality inspections. Mission Managers utilize data from these inspections to make decisions relating to safety, manpower, and funding. The overall impact of quality data on NSTS mission success was the prime reason for selecting the Quality Assurance function as a process improvement project. After a review of new technologies, it appeared that Pen-Based computers would meet improvement requirements.

In 1992, NASA funded a project to research and assess Pen-Based computer technology. The project's goal was to determine if Pen-Based computers could improve the data collection process at KSC and simultaneously reduce the use of paper forms required by the Quality Assurance Program. Before the project could begin, a thorough review of the data collection process was required.

The function of KSC's Quality Assurance Program is to collect information on the processing activities of the Space Shuttle in order to eliminate obstacles and to provide guidance to achieve maximum utilization, efficiency, and effectiveness of the processing resources [1]. In order to perform this function, data must be collected and analyzed. This is accomplished by the KSC Quality Assurance Surveillance Program. The program assigns a Quality Assurance Specialist (QAS) to an area to survey activities. The task of the QAS is to perform scheduled and unscheduled surveillance of contractor activities, and to verify, monitor, witness, and perform inspections. This surveillance often requires the use of Work Authorization Documents (WADs) which identify Mandatory Inspection Points (MIPs). If surveillance is performed against a WAD, a status stamp is required for the steps surveyed. Once the surveillance is complete, the QAS records the results of the activities on a Quality Surveillance Record (QSR) (Figure 2). At the end of the shift, all the QSR forms are collected and reviewed by the Functional Supervisor. The QSRs are then routed to the Branch Supervisor. Once the Branch Supervisor has reviewed all the QSRs, they are forwarded to Quality Engineering (QE) for review and sign-off. The QSRs are then routed to a data entry clerk for processing. The QSRs are entered into a database, compiled once a month and then transferred to a floppy disk. The floppy disk and the original paper forms are routed to the NASA Trend Analysis Group for analysis. The trending group then processes the data and produces numerous trending reports for distribution.

NO. 03707

[illegible]

Figure 2 QSR Form

After completion of the process analysis, a technology survey was performed on Pen-Based Computers and related equipment. The survey addressed the areas of portability, durability, ease of use, ability to transfer and receive data, programmability, and cost. The search turned up over 195 commercial vendors of Pen-Based computers and associated hardware and software. These computers range from 286SXs without hard drives, to 486s with over 200 MB hard drives. Pen-Based computers are classified in four distinct categories: Tablet, Pentop, Tethered Tablet, and Palmtop. The Tablet Pen Computer is about the size of a standard notebook and uses the pen as the primary input device. The Pentop Pen Computer has a tilt up screen and provides both stylus and keyboard inputs. The Tethered Pen Computer has the pen input device attached to a desktop computer. This type of pen computer is very popular with secretaries and software developers because it can tie directly into their workstation. The Palmtop Computer does not have a built-in harddrive, therefore it is lightweight and can be hand-held. The Palmtop Computer uses the pen as the primary input device and most have communication networking capabilities. The survey concluded that Pen-Based computers would be extremely beneficial in streamlining the data collection process at KSC.

SYSTEM DESIGN

System Automation Requirements:

The system requirements for the Quality Assurance data collection efforts were evaluated and narrowed down to three items: size, weight, and the ability to be tethered. These criteria are important due to the working conditions that will be encountered. The system must be light enough to carry around for an entire eight hour shift. In addition to being light, it must also be small enough so that it can be maneuvered in tight areas. The most important requirement is the capability of tethering the system to the individual using the computer. NASA Safety requires that all items used in the proximity of the Shuttle be tethered, thus eliminating the chance of damaging the Shuttle or harming others. Based on these three criteria, the PalmPad computer, a member of the Palmtop category, was selected.

System Description:

The system selected for the data collection automation is comprised of three subsystems. They are the PalmPad computer, an operating system, and a Touch Memory peripheral. This section will describe each subsystem.

PalmPad:

The PalmPad weighs 2.9 lbs., provides hand-straps, and is equipped with a 2 MB Flash Card. The Flash Card, also known as a PCMCIA card, is a removable storage device the size of a credit card. The card inserts into a slot similar to a disk drive and replaces the permanent harddrive. There are two types of PCMCIA cards: memory cards and I/O cards. The memory cards are divided into four categories: static RAM, dynamic RAM, EPROM, and EEPROM (flash memory). The memory cards are used in place of a disk drive or harddrive and range from 2 MB to 20 MB of memory. Two important advantages for using this form of storage are increased access time (about 10 times faster than a hard disk), and decreased power consumption [2]. The primary disadvantage of memory cards is the cost, approximately \$1000.00 for a 20MB card. The I/O cards, however, are used to support peripheral devices, such as modems, FAX cards, LAN adapter cards and bus adapter cards.

Operating System:

The operating systems of two leading competitors, Go Corporation and Microsoft Corporation, were evaluated for use. A closer look into each company's operating systems revealed distinct differences. One noted difference was that Microsoft's Pen for Windows did not require file or data conversion when dealing with existing applications which was a major factor in selecting their operating system. Microsoft also provided the ability to connect to a Local Area Network (LAN) which was an important feature. Pen for Windows includes all the networking solutions that are associated with the traditional Windows. The final issue, and most important, was handwriting recognition that also incorporated a handwriting learning capability. First time users complete a training exercise that requires them to print the letters of the alphabet three times. Information is then stored in that individual's own dictionary, which will be referenced anytime that individual logs onto the system. This becomes very helpful when a persistent problem recognizing a character arises. If this occurs, the user can call up the Trainer, reference the problem character, and select the correct translation. This character will then be added to the user's personal dictionary as a reference. The versatility of the above mentioned Pen for Windows features was the prime reason for selecting this software as the operating system.

Touch Memory:

Touch Memory is a peripheral device that can be integrated with the PalmPad for an electronic sign-off capability. When dealing with the legalities of electronic sign-off of forms, there are two solutions. The first solution is the capturing of a signature by use of bit map imaging. The second solution is that of a new technology called Touch Memory. Touch Memory is a innovative, nonvolatile memory chip packaged in a rugged, stainless steel can (Figure 3). This new, coin-shaped device makes it easy to transfer information simply with a touch. The reading and writing of the Touch Memory button is accomplished by momentarily contacting a Touch Button reader. Unlike the read-only bar code system, Touch Memory has a read/write capability and may replace bar-codes in the future. Touch Memory also records over 100 times the data of bar codes and is available with memory capacities of 4 KBs and more. This technology will be interfaced with the Pen-Based computers for handling signatures on the QSR forms. Touch Memory buttons will be assigned to each individual required to sign QSR forms. Sign-off will be accomplished by simply touching the button to a reader on the Pen-Based computer. This procedure transfers the individual's ID number onto the form.

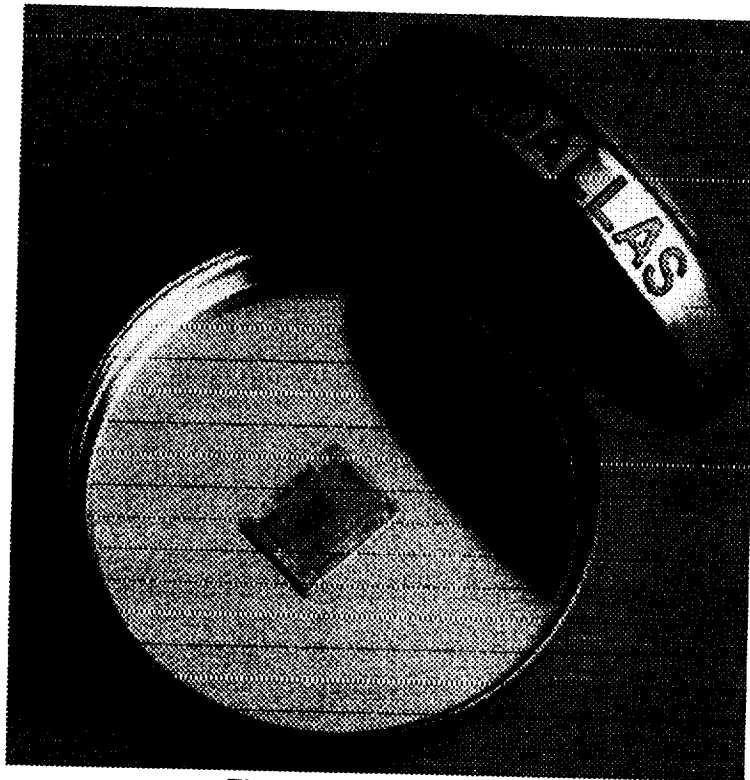


Figure 3 Touch Memory

Figure 4 QSR Entry Screen

IMPLEMENTATION

After the identification, selection and procurement of hardware and software, the system was configured to meet the unique KSC data collection program requirements. The system required a log-on, log-off capability, and data entry capability which is accomplished through the use of a Main Menu and Data Entry screens. The original paper form was altered to allow for a smooth transition throughout the form and to allow ample room for data entry. The QAS using the Pen-Based computer will first log onto the system. The Main Menu will then appear. The "Add QSR" button will be selected which will bring up the entry screen as shown in Figure 4. The entry screen has been designed with legal value tables, and required field checks. Legal value tables have been set up for location, vehicle, organization, QAS, work area, program, contractor, system, failure code, cause, responsible organization, and corrective action fields. Certain areas of the form are automatically updated when a key value is entered. For example, when the name of the QAS is entered, the supervisor name, the QAS's mail code and shift will automatically be filled in. After each QSR record is completed, a verification check will be automatically performed. The QAS proceeds through the form, filling in the appropriate information, until all data entry is completed. At this point, an option for adding a new QSR or exiting the system is provided. Regardless of the action selected, a check will be made to verify that all required fields have been filled in. If a required fields has been left blank, an error message will appear and no further action will be permitted until the error condition has been corrected. The QAS is then allowed to add a new QSR or exit the system using the touch memory button.

The complete QSR data collection system consists of a handheld Pen-Based computer utilizing Pen for Windows as the operating system. This system collects and stores data in a flat ASCII format on the PCMCIA card. At the end of each shift, the QAS will extract the PCMCIA card and place it in an external drive that is located at a stationary workstation. This data will then be uploaded to the network, and appended to the centralized database which stores all QSR data.

Testing and Validation:

System testing and validation was accomplished over a four month period by running the Pen-Based computers in parallel with the paper forms. The data from both systems was compared to make sure that it had not been corrupted or lost.

Training:

Pen-Based computers are extremely user-friendly and requires minimal training. System training at KSC will be conducted in two phases: Pen-Based unit familiarization and forms entry. It is designed to provide QAS personnel with "hands-on" experience, and will be required prior to certification.

CONCLUSION

Pen-Based computers are extremely versatile and can be used to automate virtually any data collection process. They can be tailored to unique requirements, integrated with any computer system, and networked for data exchange. Their use at KSC for the data collection, transfer, and storage process has greatly improved management decision support. Delays in shuttle processing related to data deficiencies have been reduced significantly. Pen-Based computers have been extremely useful in streamlining the NASA Quality Data Program and will be further utilized in other paper processes at KSC and other NASA centers..

ACKNOWLEDGEMENTS

The data collection, Pen-Based computer system has been developed with the funding from NASA (contract NAS10-11675).

REFERENCES

1. National Aeronautics and Space Administration, Quality Assurance Procedure #11, Kennedy Space Center, FL, 1991.
2. Bruce Schneier, "Inside the PCMCIA storage standard," MacWeek, January 11, 1993, p. 50.
3. Deputy Director Gene Thomas, "Recycling at KSC," Spaceport News, January 15, 1993, p. 3.
4. Frederic E. Davis, "Will you be running your next computer with a pen?," PC Week, November 25, 1991, p. 118.

"THE VERTICAL"

Stephen L. Albert
Ergonomic-Interface Keyboard Systems, Inc.
P.O. Box 2636
La Jolla, CA 92038

Jeffrey B. Spencer
Ergonomic-Interface Keyboard Systems, Inc.
P.O. Box 2636
La Jolla, CA 92038

ABSTRACT

"THE VERTICAL" computer keyboard is designed to address critical factors which contribute to Repetitive Motion Injuries (RMI) (including Carpal Tunnel Syndrome) in association with computer keyboard usage. This keyboard splits the standard QWERTY design into two halves and positions each half 90 degrees from the desk. In order to access a computer correctly, "THE VERTICAL" requires users to position their bodies in optimal alignment with the keyboard. The orthopaedically neutral forearm position (with hands palms-in and thumbs-up) reduces nerve compression in the forearm. The vertically arranged keypad halves ameliorate onset occurrence of keyboard-associated RMI. By utilizing visually-reference mirrored mylar surfaces adjustable to the user's eye, the user is able to readily reference any key indicia (reversed) just as they would on a conventional keyboard. Transverse adjustability substantially reduces cumulative musculoskeletal discomfort in the shoulders. "THE VERTICAL" eliminates the need for an exterior mouse by offering a convenient finger-accessible cursor control while the hands remain in the vertically neutral position. The potential commercial application for "THE VERTICAL" is enormous since the product can affect every person who uses a computer anywhere in the world. Employers and their insurance carriers are spending hundreds of millions of dollars per year as a result of RMI. This keyboard will reduce the risk.

ERGONOMICS AND RMI/CTD

The understanding of ergonomic factors for computer workstations is a relatively new area of research and application. The word 'ergonomic' explains the interaction of people to their environment and only first appeared in the workplace in the early 1980s with regard to the meat-packing industry. Although the need for ergonomic office furniture and furnishings has been recognized since 1986, due to the rapid increase of Repetitive Motion Injuries (RMI) and Cumulative Trauma Disorder (CTD) at computer workstations, modifications have been slow in coming. This is due to the perceived high-cost factors to the employer and the absence of governmental regulations.

During 1991 and 1992, media sources focused on business owners who had a difficult time paying for rising insurance costs due to increased workers' compensation claims for non-accident related injuries. The media began concentrating its attention on ergonomic solutions to work environment problems other than the computer keyboard. They blamed the computer industry, the video display terminal and lag time by a majority of employers, who over the past decade have used technology to double production, but have failed to protect the worker using that technology. By mid-1992, full media attention was focused on the recognition that the traditional computer keyboard design was the primary cause of repetitive motion injuries for computer keyboard users.

Repetitive motions (keystrokes) are an essential part of a computer operator's work function. According to the Bureau of Labor Statistics, repetitive motion injury claims, such as Carpal Tunnel Syndrome, have grown at an alarming rate over the last decade. These injuries can cost employers \$30,000-\$80,000 in health insurance, sick pay, disability, and workers' compensation benefits per incident.

CARPAL TUNNEL SYNDROME

The major RMI problem with which most people are familiar is Carpal Tunnel Syndrome (CTS). CTS is caused by repetitive, forceful, quick and uninterrupted tasks common to computer keyboard operators. The carpal tunnel, which runs through each wrist into the hands, houses the median nerve and nine tendons which control movement of the thumb, forefinger, middle finger and half the ring finger. The tunnel is formed by the carpal bones on the back of the hand and the transverse carpal ligament on the palm. Repetitive keystroking can cause swelling around the tendons, which puts pressure on the median nerve, causing pain and reducing hand function. The physical condition of the keyboard operator may accelerate CTS and quicken the onslaught of pain. This is especially true for many women who, during pregnancy, hold on to body fluids and are prone to bloating at the wrists. When overdone, other outside activities such as tennis and knitting may also accelerate CTS.

Pronation, the twisting of one's wrist so the palms are facing downward, while working on the keyboard, has been identified by medical and orthopaedic specialists around the world as a major cause of Repetitive Motion Injuries. Employees working at computer workstations are potential victims of RMI. The immediate focus are the individuals who work on computer keyboards for extended periods of time each day. Workers in this endangered class are employed by federal, state and local governments, major corporations, large industries and major service organizations which interact with other public and private companies. Their jobs may consist of six to eight hours of data input per day without proper breaks for exercise and stretching. Many orthopaedic physicians recognize the severity of the problem, and have developed specific exercise programs which may help combat Carpal Tunnel Syndrome for computer operators. Unfortunately, these are only band-aids to the real problem of improper hand and body positioning.

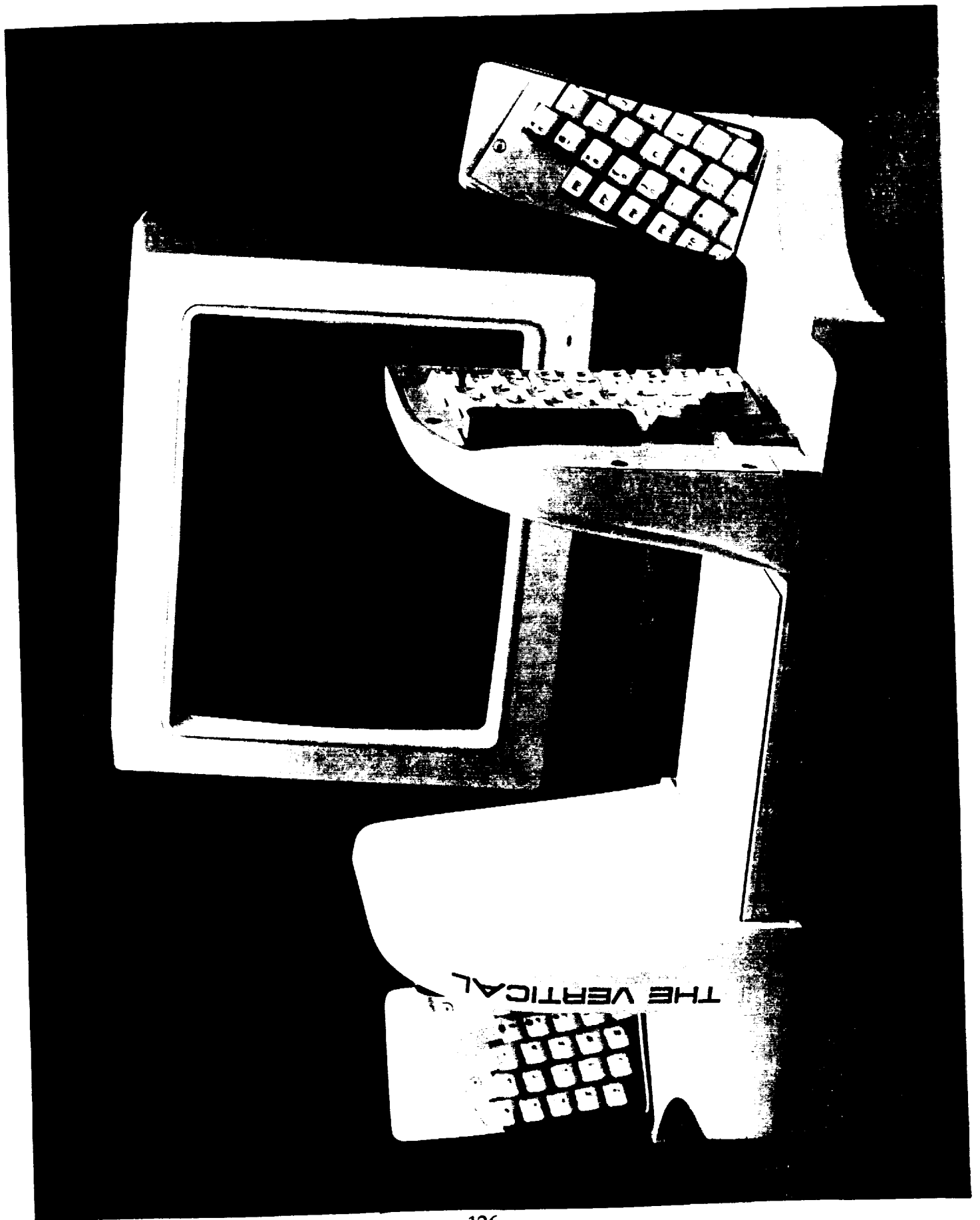
COMPUTER COMPANY EMPHASIS

Computer companies and manufacturers have always placed their primary focus on the production of faster and more reliable computers and ignored the physical needs of computer operators. They have produced inexpensive keyboards which, instead of being "value-added", are considered "commodities" which unfortunately must be included to operate the computers. Even in the sight of emerging class action lawsuits (charging computer and keyboard companies with negligence in continuing to manufacture products detrimental to users), the companies have ignored the evidence and continued to manufacture ill-designed keyboards.

It must be realized the large percentage of office workers who have developed RMI from computer keyboard usage did not grow up with computers. RMI occurred over the last 10 - 15 years due to the extensive use of computers in the work environment. However, children are growing up using computers from a very early age. If the keyboard design is not corrected, RMI will affect many more people in epidemic proportion at a much earlier time in life.

"THE VERTICAL"

"THE VERTICAL" keyboard, designed and developed by Ergonomic-Interface Keyboard Systems, Inc., in La Jolla, California, is patent protected in the United States and patent pending internationally. Its design has received endorsements from hand surgeons and clinical orthopaedic therapists as well as from OrthoMed, the Hand Rehabilitation Center of the University of California, San Diego Medical Center. "THE VERTICAL" is the only ergonomic keyboard which UCSD has approved for testing with its patients.





Beyond the constant barrage of media attention regarding RMI such as Carpal Tunnel Syndrome, the demand for "THE VERTICAL" is evident from the scores of letters, phone calls and faxes from around the world received by The Company since the first few magazine and newspaper articles about it appeared in October of 1992. Many inquiries are from safety engineers and ergonomic experts representing large corporations who understand the problem in the industry and view "THE VERTICAL" as a viable solution. Inquiries from federal, state and local governmental agencies have been received and are awaiting production of the final manufactured units. International companies have requested information on terms the Company would require for them to become manufacturing and distribution arms.

THE MARKET

No product could have a more ready-made market than "THE VERTICAL". RMI is a major world-wide health problem. Employers, and their insurance carriers, are spending literally hundreds of millions of dollars per year as a result of RMI. The product is the most eagerly sought item in the computer-related industry; a keyboard which preliminary judgment leads us to believe should substantially reduce the risk of RMI (including Carpal Tunnel Syndrome).

Based on Standard & Poors October 1991 Industry Surveys, not including keyboards from Apple, Omega, Tandy, etc., there are seventy-five million IBM and compatible keyboards in the United States. With well over one hundred million (100,000,000) keyboards in use in the U.S. market alone, each one utilizing an ill-designed keyboard, the market for a "value-added" keyboard is obvious. Penetrating fifty percent (50%) of the existing market would mean selling over fifty million (50,000,000) keyboards in the United States alone. Conceptually however, the market area for a correctly designed ergonomic computer keyboard is throughout the world, and the U.S. portion of that is under twenty percent (20%).

ORTHOPAEDICALLY OPTIMAL KEYBOARD FACTORS

The words "ERGONOMIC KEYBOARD" have been used to identify a few new keyboard designs recently developed by individuals who are not employed by computer and keyboard companies. However, those words do not determine the degree of healthfulness for the user. Our extensive research and work with orthopaedic surgeons and clinical rehabilitation therapists has shown the optimal progressive ergonomic keyboard:

- Encourages optimal seating positioning (MOST IMPORTANT)
- Eliminates pronation (downward rotation of the hand and forearm)
- Eliminates arm/shoulder flexion (extending upper arm forward from shoulder)
- Adjusts to the user's torso width
- Reduces transitional motion (removing hand from keyboard to use a mouse)
- Accommodates the shape and movements of user's hands,
- Increases productivity / decreases fatigue

"THE VERTICAL" is designed to address critical factors causing RMI in association with computer keyboard usage. In order to access a computer correctly, "THE VERTICAL" requires users to position their bodies in optimal alignment with the keyboard. The orthopaedically neutral forearm position (with hands palms-in and thumbs-up) eliminates hand and forearm pronation, which reduces nerve compression in the forearm. The vertically arranged keypad halves serve to ameliorate onset occurrence of keyboard-associated RMI by eliminating the extremes of wrist flexion (up and down movement of the wrist), shoulder and arm extension and ulnar deviation (outward rotation of the wrist). By utilizing visual-reference mirrored mylar surfaces adjustable to the user's eye, the user is able to readily reference any key indicia (reversed) just as they would on a conventional keyboard. Transverse adjustability (movement like an accordion to a locked position) to user's torso width substantially reduces cumulative musculoskeletal discomfort in the shoulders while reducing wrist deviation. "THE VERTICAL" eliminates the need for an exterior mouse by offering a convenient finger-accessible cursor control, while the hands remain in the vertically neutral position resulting in increased productivity through less transitional motion.

Patients who have experienced discomfort on traditional computer keyboards have found relief when using "THE VERTICAL". Further research with keyboard users showed individuals, and especially industry, will not accept a product which forces them to relearn the keyboard or takes extensive amounts of time to regain normal speed and accuracy. Based on the comments of hundreds of people who have interfaced with "THE VERTICAL" prototype, an average touch typist will regain speed and accuracy within hours.

COMPETITION

"THE VERTICAL" will compete with other companies which develop and market ergonomic computer keyboards, some of which are of greater size and may have greater financial resources. Ergonomic-Interface Keyboard System Inc.'s major competitors are the manufacturers of new "ergonomic" keyboard styles. However, those new keyboard styles which allow the user to pronate their hands and wrists during keystroking, maintain a "flat" keyboard similar to the major manufacturers, thereby incurring the same potential problems to the user. Others stray from the standard QWERTY key configuration and eliminate traditional typing formats altogether. They insist the user learn a new operational language in order to operate the board. Since, over the years, other unconventional keyboards such as the Dvorak System have been introduced yet have not been accepted by industry due to the need of retraining the user, it is felt by industry experts any keyboard which strays from the QWERTY format will not be accepted by the masses. One of the competitors states the optimal, ergonomic keyboard operates with the user's hands in a vertically oriented position, yet their keyboard does not offer the user that possibility without eliminating the visual connection from the user to the board. Existing major computer keyboard manufacturers and major computer companies which manufacture their own keyboards are not considered competition due to existing lawsuits against them stemming from their ill-designed "flat" or "conventional" keyboards.

Competition exists between the Company and some ergonomic experts who ignore the physical problems which the computer keyboard has been accused of creating. Their opinion is with proper body positioning and the correct use of properly designed ergonomic furniture, the problems of RMI and CTD could be eliminated. This is not the opinion of most medically and/or orthopaedically trained physicians. It does not address the damage keystroking plays on the user's body while the hands and wrists are pronated and arms are projected forward from the shoulders. Body position and furniture type alone cannot correct RMI and CTD problems. However, the Company does agree proper body positioning and the correct use of properly designed ergonomic furniture should be utilized while using "THE VERTICAL" so as to enhance the value of the keyboard's design.

Some experimentation has begun with hand written data entry, but this has limited application at best (for limited use by physically disabled operators, etc.). The error factor is also a serious problem with hand written data entry.

FUTURE PRODUCTS

The long term goals of The Company are towards the modification of other data input devices to obtain benefits similar to those of "THE VERTICAL". Adaption of the basic design to meet specialized keyboard styles for governmental and research usage is expected to be an easy transition. Another logical extension of this design, which is now being worked on by The Company, is a vertically oriented Stenograph device.

A SYSTEMS APPROACH TO COMPUTER-BASED TRAINING

Gaylen W. Drapé
 ENSCO, Inc.
 445 Pineda Court
 Melbourne, FL 32940

ABSTRACT

This paper describes the hardware and software systems approach used in the Automated Recertification Training System (ARTS), a Phase II Small Business Innovation Research (SBIR) project for NASA Kennedy Space Center (KSC). The goal of this project is to optimize recertification training of technicians who process the Space Shuttle before launch by providing computer-based training courseware. The objectives of ARTS are to implement more effective CBT applications identified through a needs assessment process and to provide an enhanced courseware production system. The system's capabilities are demonstrated by using five different pilot applications to convert existing classroom courses into interactive courseware. When the system is fully implemented at NASA/KSC, trainee job performance will improve and the cost of courseware development will be lower. Commercialization of the technology developed as part of this SBIR project is planned for Phase III. Anticipated spin-off products include custom courseware for technical skills training and courseware production software for use by corporate training organizations of aerospace and other industrial companies.

INTRODUCTION

Global competition and rapid changes in technology have increased the demand for employee education and job training. There are three major reasons for this increase in demand [1], the first being that displaced workers need to be retrained. It is estimated that most workers will change jobs five or six times during their lives. Due to the dynamic nature of the U.S. economy, 1.5 million workers are permanently displaced each year and require assistance to reenter the workforce. By the year 2000, it is estimated that 5 to 15 million manufacturing jobs will require different skills than for today's jobs, while an equal number of service jobs will become obsolete.

Second, the work being performed at most companies is becoming increasingly complex. The use of computers and more sophisticated business processes require many employees to relearn how to perform their jobs. Furthermore, competition demands that there be constant change in the products and services that companies offer. This causes a ripple effect throughout the company in bringing these new products and services to customers.

Changes in the organizational structure of companies also increase the demand for job training. One result of the difficult economic climate is that many companies are downsizing, increasing the need for cross training of workers. One of the principles of Total Quality Management (TQM) is the expansion of employee empowerment, with teams of employees performing a function or process with little or no direction from traditional management. Companies that are turning to TQM principles are finding that employees are unqualified for this empowerment without a large investment in training.

The United States' educational system fails to prepare many employees for the challenges of the modern workplace. The poor performance of U.S. high school graduates relative to their foreign counterparts on standardized tests is well publicized. Without supplementary training, the level of education of available American workers frequently fails to meet the requirements of employers. A recent study shows that one-fifth of displaced workers lack a high school education and that 20 to 40 percent of these workers are considered functionally illiterate.

Due to changes in the economy and increases in skill level requirements for the workforce, many companies have expanded their employee training programs. Typical employee development programs at many large companies now include remedial training in reading, writing, and basic mathematics. Also, increased quality and safety requirements have caused companies to institute formal job or skill certification programs. However, the increased necessity of workforce training is costly in terms of time away from the job, travel costs to and from training sites, and expenses associated with classroom facilities, instructors' salaries, and administration.

Means must be found to make employee training more cost effective. Computer-based training (CBT) has been available for many years as either an alternative or a supplement to classroom training. Government-sponsored research and companies' experience have shown CBT has the following potential benefits over conventional instructor-led training.

- Training can be delivered and administered at lower cost.

This is possible because a computer can be placed at or near the work site at a time that is convenient to the learner. Traditional classroom instruction settings and instructors are unable to provide the level of flexibility in location and time that a computer is able to. CBT can be delivered close to the work site and can be scheduled at a time that is more convenient to the learner. Delivery of training on the computer also makes the course content more consistent and maintainable.

- Training can be accomplished in less time.

CBT allows individualization of lessons, whereby students access only the information that they need to learn. The modern "point-and-click" user interface commonly used in CBT gives the learner increased control of the lesson and allows the training to be self-paced. Thus, faster learners are able to complete lessons in less time.

- Computer-based training improves learning when the training program is effectively developed.

By building the capability for increased interaction into the lesson, immediate feedback and remediation is available to students. It is now possible to integrate audio and video into CBT presentations that appeal to multiple senses and various learning styles of students. Therefore, learners are more highly motivated to complete the training program.

- Training achievement can be measured and tracked more easily.

CBT enables tests and assessments to be embedded into the course, making recordkeeping easier. Student performance data can be stored for later analysis to improve lessons.

Despite the potential benefits of CBT, the technology has not yet fulfilled its promise to make significant improvements in employee education and job training. Most applications of CBT are used in the United States military for combat training of personnel. Many of the training programs created for the military used specialized high-end applications developed for proprietary computer hardware. CBT did not become widely used in industry because it was too expensive to develop and the need for sophisticated training was not justified. Within the past few years, however, advances in personal computer hardware and software have made CBT more capable and affordable. The remaining challenge lies in demonstrating that CBT can be a cost-effective method of employee training in commercial industry.

PROJECT BACKGROUND

In January 1991, ENSCO, Inc. in Melbourne, Florida was awarded a Small Business Innovation Research (SBIR) contract to provide the Kennedy Space Center (KSC) with enhancements to its computer-based training. The focus of our project was to develop courseware for recertification training of technicians who perform pre-launch processing of the Space Shuttle. Processing of the shuttle vehicles requires that shuttle technicians be certified in approximately 500 technical skills. These skills include the operation of specialized test equipment, as well as performance of various types of mechanical and electrical repairs to the shuttle vehicle and its major systems. KSC's technical training program requires that most technicians be recertified annually by attending a classroom refresher course and taking an examination. With each technician holding approximately seven certifications, this process takes more than 6000 worker hours per month away from work schedule. The purpose of ENSCO's SBIR project is to apply CBT technology to existing recertification training courses, the result being the more effective delivery of training in less time than with conventional classroom training.

Investigations of the target population and the types of work performed produced a number of factors that support the benefits of CBT over instructor-led training for skill recertification. First, shuttle technicians have varying training needs that tend to favor self-paced learning. Training requirements differ for electrical technicians,

mechanical technicians, and quality control inspectors. Depending on the individual technician's job, a particular knowledge or skill may be performed at a varying time interval. This variance can result in a heavy "forgetting curve" of the skill for some technicians but in little or none for others. Evidence also indicates that, due to differences in their ages and reading comprehension levels, technicians use diverse learning techniques.

Another major factor favoring CBT is the dynamic nature of the shuttle processing work schedule, which depends upon the particular operations and maintenance tasks performed on the shuttle vehicle after each mission. Coordinating technicians' training requirements with the changing work schedule presents challenges to the existing certification program. There are a number of instances in which a job cannot be performed as scheduled because a technician cannot attend the recertification class. There are also occasions when too many "no shows" in the classroom cause the cancellation of a scheduled training course.

In January 1991, ENSCO began the research and development (R&D) effort on the process to convert existing recertification courses to CBT. This project required a strong working relationship between ENSCO and the Lockheed Space Operations Company (LSOC), which was the prime contractor responsible for shuttle processing and implementation of the technician training program. ENSCO also enlisted the services of the University of Central Florida to consult on training systems analysis and design.

During Phase I of the SBIR contract, an early prototype application was delivered on a stand-alone PC. This prototype demonstrated the feasibility of CBT for recertification of typical shuttle processing tasks and showed that CBT could be accepted by technicians. However, Phase I research also showed the development process for sophisticated CBT was complex, particularly where multimedia computer technology was required. It was also determined that CBT would eventually need to be delivered by way of a network to achieve the greatest usage cost savings. Therefore, improvements in both the development and distribution systems were needed for CBT to be produced and distributed on a larger scale.

ENSCO's goal during Phase II of the project was to apply recent commercially available technology to the CBT process, which would result in a more efficient and cost-effective medium for recertification training. Our technical approach consisted of two major objectives:

1. To provide a process for selecting good applications for CBT and determining the content of the CBT programs.
2. To make the system for developing and maintaining CBT programs less costly and easier to use.

Each of these objectives was addressed to overcome weaknesses in the current CBT system. In the following sections, these objectives are discussed in more detail.

NEEDS ASSESSMENT PROCESS

The first objective was satisfied with the needs assessment process. Needs assessment is a term that has many meanings in training and human performance literature. [2] The definition that seems most applicable to this project is *the analysis of the training situation for the purpose of defining the requirements of a CBT development project*. The needs assessment is conducted prior to the design of the CBT program. Its purposes are to identify the training goals and objectives and to select the appropriate media with which to present the content. A thorough needs assessment ensures that the instructional objectives are congruent to the performance job or task. A needs assessment also provides the basis with which to make rational decisions about how to apply CBT in the training program.

The needs assessment process applied by ENSCO can be described in two stages, as shown in Figure 1. The first stage is task analysis, which determines the training objectives – what the trainee needs to know or needs to do to perform a task. Prior to analysis, data are collected from the existing lesson plans, from interviews with subject matter experts and, if necessary, from observations of the task being performed. These data are then analyzed and a list of the knowledge and skills needed to perform the task is compiled. It is also important to identify prerequisite skills and knowledge during the analysis to avoid expending effort on developing unnecessary training content.

The second stage of the needs assessment process, media selection, determines the most effective way to provide the training. The inputs to the media selection process are the target audience characteristics, the training objectives, and the environment in which the training program will be deployed. The actual selection process ranges from a simple checklist verifying that CBT is a viable alternative to a comprehensive model for determining the best media to fit a particular set of training requirements and constraints. Possible choices include a hypertext document, an interactive multimedia courseware program, a system simulation, or an intelligent tutoring system.

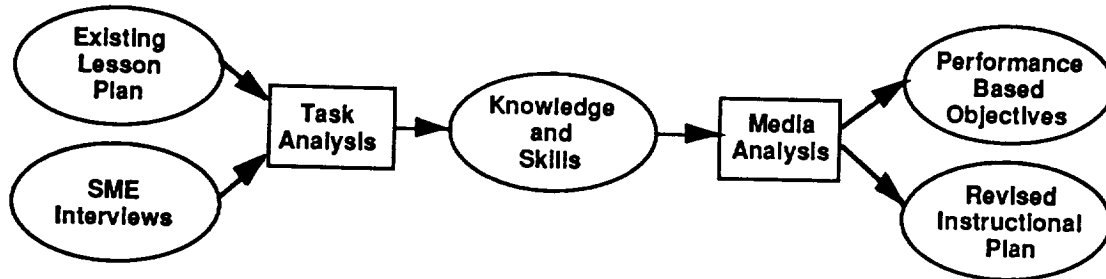


Figure 1. Needs Assessment Process.

The outcomes of the needs assessment are a data base of training objectives and a revised plan of instruction. The objectives data base usually has a hierarchical structure in which broad job performance goals, also known as terminal objectives, are divided into specific learning (or enabling) objectives. Building the objective hierarchy is a crucial step in CBT development, as it becomes the foundation for the courseware design. Each objective should be linked to storyboards for laying out the instructional content and to the testing items for verifying that the student has met each objective. The second outcome of the needs assessment is a plan of instruction. This plan should include the level of the media chosen to fit the target audience, the training environment, and the total cost estimated for the project.

There are major benefits to having a needs assessment. First, the assessment generates information with which to justify that the selection of CBT will result in improved trainee job performance. Second, there is a reduction in course design and development time as a result of the assessment's identification of specific performance-based objectives for the CBT program. This results in the training content being more relevant, increasing the probability that students will accept and use the program.

CBT PRODUCTION SYSTEM

One of the reasons that CBT is not widely used is that good CBT courseware has high development costs. Producing effective CBT is a complex undertaking that requires a team whose members have both creative and technical skills. These skills include:

- script writing
- graphic arts
- audiovisual production
- subject matter expertise
- instructional design
- computer programming
- systems engineering
- performance measurement and assessment

Building interactive CBT programs that are instructionally sound and interesting to the student requires a structured, yet flexible, design approach. CBT should be designed and built according to guidelines for screen design, instructional strategies, and testing methods. These guidelines have been established by educational psychologists and human factors engineers. For this project, the detailed requirements and standards used in the CBT lessons were established during the needs assessment phase of the project.

A CBT production system was created and was used by development team members to build courseware according to established standards and guidelines. The hardware and software components of the production system

were configured using commercially available computer technology. Three major technical concerns influenced the creation of the production system:

1. The hardware and software needed to be compatible with those used at KSC.
2. The most sophisticated courseware applications to be produced would require interactive videodisc and digital audio capability.
3. Some members of the CBT development team who would be using the system lacked extensive programming expertise.

Since most of the CBT at KSC uses the IBM PS/2 platform, we selected compatible CBT production system hardware. This hardware configuration consists of the following components, as shown in Figure 2:

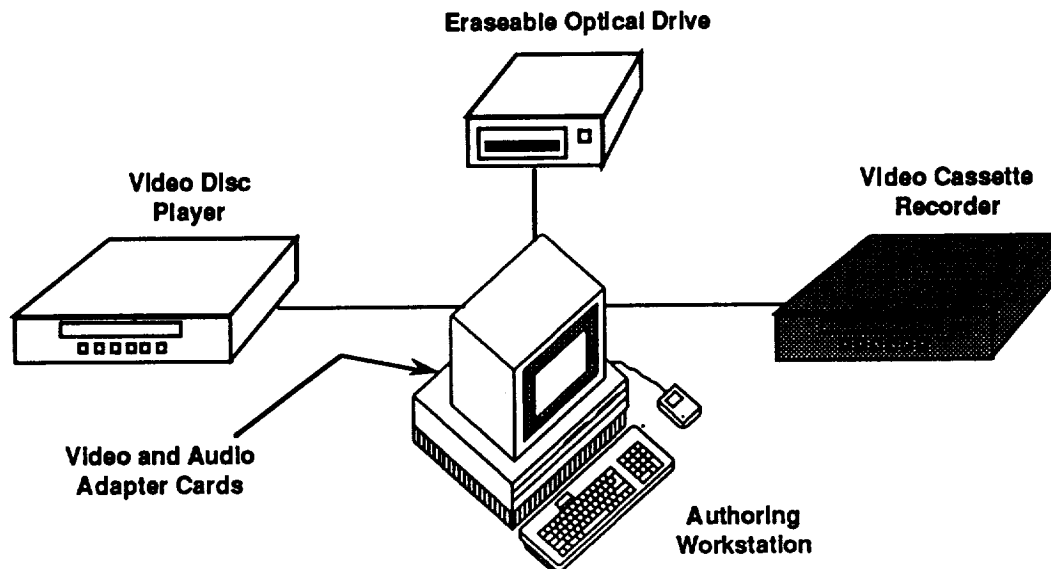


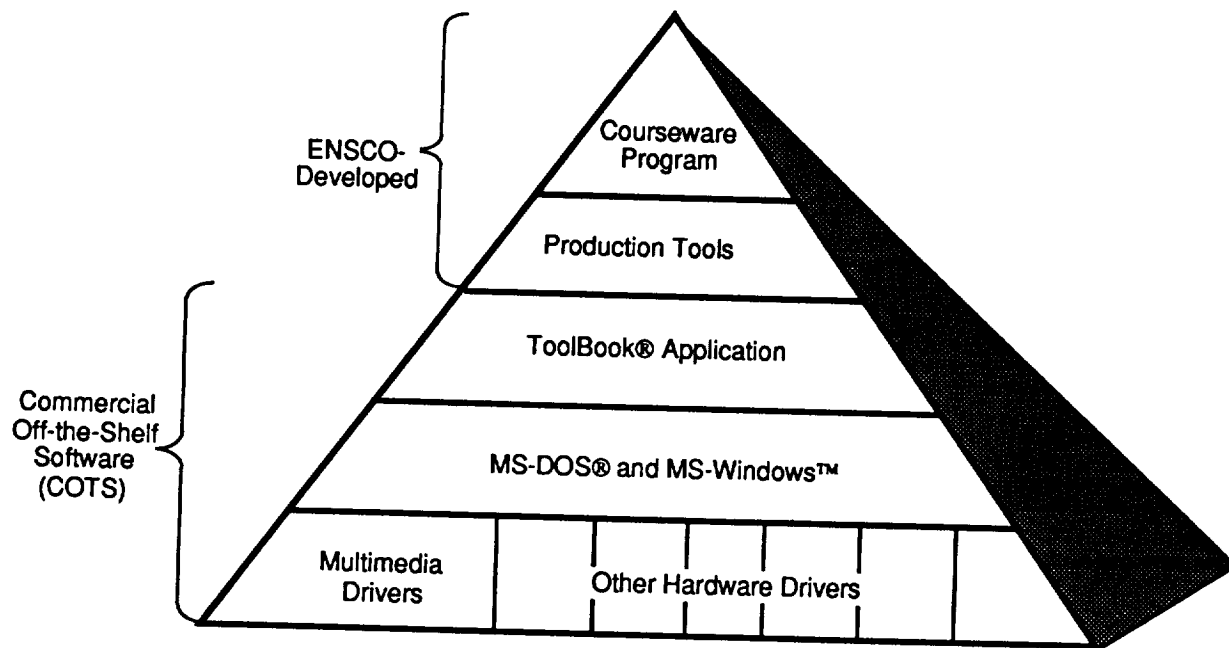
Figure 2. CBT Production System Hardware Configuration.

Most of the effort in creating the production system was expended in building the software environment. The system needed the capability for allowing the production of sophisticated CBT programs, while providing ease of use for members of the development team who were not experienced programmers. The software approach taken by ENSCO used a commercial-off-the-shelf authoring package as a foundation or shell and to integrate customized CBT production tools within the shell. The software configuration can be described as a pyramid, which is illustrated in Figure 3. Using this system, courseware programs could be written using the capabilities of the authoring package, the operating system, and the other system software that allowed operation of peripheral devices such as the audio card and videodisc player. The authoring package, Asymetrix® ToolBook®, was chosen because of its compatibility with other CBT development efforts at KSC. This version of ToolBook® runs under Microsoft Windows™ and comes with multimedia libraries that contain software drivers for audio and video hardware devices.

The major features of the production system environment are listed below.

- Password-protected access for student, developer, and system administrator
- Hierarchical structure of scripting programs for ease of maintenance
- Standard logic flow for different modes of instruction and reference, ensuring all training objectives are achieved by the student

- Templates for screen construction that assume common user navigation and media controls
- High level language for creation of time-sequenced presentation of media on the screen



MS-DOS® is a registered trademark of Microsoft Corporation.
 Windows™ is a trademark of Microsoft Corporation.
 ToolBook® is a registered trademark of of Asymetrix Corporation.

Figure 3. CBT Production Software Environment.

- Test engine for construction of test questions with question-answer randomizing capabilities and feedback
- Development tools for showing and hiding media objects and for debugging the programs

The benefits of the production environment include quality and functionality consistent with instructional system design principles, reduction in programming expertise required, and provision of a facility for program reuse and maintenance.

PROJECT RESULTS

Since Phase II of the project began in May of 1992, the system and technical approaches have been evaluated with actual KSC training courses that were converted to CBT applications. These pilot applications were selected primarily to illustrate how CBT could be used for training over a wide range of technical subject matter. The applications were also evaluated using the needs assessment process to determine their appropriateness for CBT. Examples of the types of training applications for which CBT programs were developed included:

- Operation of a helium mass spectrometer for leak testing shuttle components
- Operation of a computer-controlled laser tool for measuring steps and gaps between Orbiter tiles
- Safety precautions to be taken to counteract the hazards involved with high-pressure systems

In the remaining five months of Phase II, we plan to deploy and evaluate the pilot courses in various training environments with diverse target populations. We designed pilot courses to use different media at different levels of sophistication – from a hypermedia reference manual to interactive multimedia courses for recertification with built-

in proficiency testing. The delivery methods by which the pilot CBT applications will be evaluated include stand-alone PCs, on-line local area networks, and concurrent training applications. CBT applications will be evaluated for use as a means of automated recertification, as a classroom training aid, and as a reference tool for the workplace.

Initial data gathered from pilot studies show that the new approach favorably influences the training process. Using self-paced CBT, we expect that students will be able to cover course content in 25 to 50 percent of the time it would take them to cover the same material in a traditional classroom setting. We also expect a significant reduction in CBT development time because less programming will be required. Finally, we have CBT design standards to follow, making courseware more consistent and maintainable.

COMMERCIAL APPLICATIONS

The plan for Phase III of this SBIR project is to commercialize the technology developed as part of this project. As explained earlier, the market for CBT training in government and industry is expected to expand because of changes in the economy and demographic trends in the workforce. Insufficient numbers of instructors available to meet increasing training demands, as well as rapid growth in computer technology, will increase CBT's market potential by its capability to disseminate information and increase interactivity with students at a steadily decreasing cost.

Our primary targets for marketing CBT technology are training departments of small and mid-size companies that require custom training and skills certification courses. A secondary marketing opportunity is to produce software to augment major CBT authoring systems.

ENSCO's systems approach is to harness technology to make CBT a more efficient and effective training method. The primary ways to make this happen are to select appropriate CBT applications and to specify lesson content accurately through the needs assessment process. Another means by which to increase the value of CBT to commercial industry is to reduce development and maintenance costs through enhancements of the CBT production system.

REFERENCES

- [1] Educating America, Bowsher, J., John Wiley and Sons, New York, NY, 1989.
- [2] The Design, Development, and Evaluation of Instructional Software, Hannafin, M., and Peck, K., Macmillan, New York, NY, 1989.

Om IT

COMPUTER SOFTWARE

AUTOMATIC TRANSLATION AMONG SPOKEN LANGUAGES

Sharon M. Walter and Lt. Kelly Costigan
Rome Laboratory/IRAA
Griffiss Air Force Base, NY 13441-4114

2500
P. 4

ABSTRACT

The Machine Aided Voice Translation (MAVT) system was developed in response to the shortage of experienced military field interrogators with both foreign language proficiency and interrogation skills. Combining speech recognition, machine translation, and speech generation technologies, the MAVT accepts an interrogator's spoken English question and translates it into spoken Spanish. The spoken Spanish response of the potential informant can then be translated into spoken English. Potential military and civilian applications for automatic spoken language translation technology are discussed in this paper.

THE MACHINE AIDED VOICE TRANSLATION (MAVT) SYSTEM

During times of military conflict it is important to acquire intelligence information quickly. The best sources of timely information, however, are often foreign-language speaking people: defectors from the opposition's camp, Prisoners of War, and civilians from the conflict area. Whenever and wherever conflict occurs, military linguists who are versed in the particular language and dialect of potential informants, familiar with the Commander's military strategy, and knowledgeable about interrogation techniques are a valuable asset --- and an extremely rare commodity.

The Machine Aided Voice Translation (MAVT) system is an early prototype demonstration of the application of current speech processing technology to help compensate for the shortage of suitably trained and experienced linguists. It allows a less skilled interrogator to "screen" potential informants. When an interrogator with little or no foreign language skills asks questions by speaking into the microphone in English, the MAVT translates the questions into machine-spoken Spanish. Upon hearing each question in his/her native language, the potential informant responds by speaking into the microphone and the system translates the response into spoken English. Based on the interrogator's perception of an informant's cooperativeness, reliability of information, and relevance of the information to the Commander's intelligence requirements, the potential informant can be referred to a more skilled interrogator for further, deeper questioning.

The MAVT display is shown in Figure 1 on the next page. The user selects 'male' or 'female' and 'English' or 'Spanish' to indicate to the system the gender of the speaker and whether he/she is an English-speaking interrogator or a Spanish-speaking interrogatee. Providing the MAVT with knowledge of the gender of the speaker allows better speech recognition due to the more appropriate use of either a male or female speech "model."

Inputs to the system must be restricted to those that use words from the system's Spanish and English dictionaries (lexicons) and those which use a word order allowed by its grammars. There are two MAVT grammars. One allows questions and answers about biographical information, so examples in English are: "What is your name?" and "Indicate your unit designation."; examples in Spanish are: "Mi rango es teniente general" and "Naci en Santa Clara."* The second grammar focuses on mission-related information such as (in English:) "Why was your unit moving out to the south?" and "Is the main force heading in that direction?", and (in Spanish:) "Proteger el puesto de comando del regimiento" and "Su misio'n es encontrar unidades americanas." The display includes a scrollable list of examples of the inputs that are accepted by the speech recognizer. When a Spanish speaker is anticipated the display lists examples of acceptable Spanish responses.

The MAVT prototype is serving as the design basis for a follow-on development that will extend the English-Spanish translation vocabulary and expand system capabilities to include English-Arabic and English-Russian spoken language translations. The follow-on system will be completed in late 1996.

Language Systems Incorporated, of Woodland Hills, California, developed the MAVT and is the primary contractor for the advanced development model.

Hardware/Software Architecture. Briefly

The speech recognition system of the MAVT is the PE200 Phonetic Engine produced by Speech Systems, Inc. (SSI). The Phonetic Engine accepts speaker-independent, continuous speech. That is, it does not have to be trained to recognize any particular voice, and users can speak quite naturally and fluidly without pausing between each word as would be required by an isolated word speech recognizer.

* English and Spanish examples are not intended to represent questions and their respective answers.

What is your duty position
 Was your mission offensive
 What was your mission
 Why was your unit moving out to the south
 Is the main force heading in that direction
 Can the forward element see our tanks from the road
 Are they repositioning to the right of your unit
 What kind of vehicles do they have
 How many tanks do you have

CHECK
 What is your birth date

Figure 1
 MAVT Display

MAVT speech output is provided by a DECTalk DTC01 speech generator from the Digital Equipment Corporation (DEC). The DECTalk accepts text output and converts it into spoken words. The pitch and speaking rate of the DECTalk may be modified.

The core, language translation software of the MAVT is LSI's DBG natural language processing system hosted on a Sun Microsystems Workstation (SPARCstation). DBG was extended for this project with a multilingual lexicon, a multilingual morphological component, and a language-independent syntactic parser. DBG works by deriving semantic (meaning) representations of inputs and translating them into the language which is to be output.

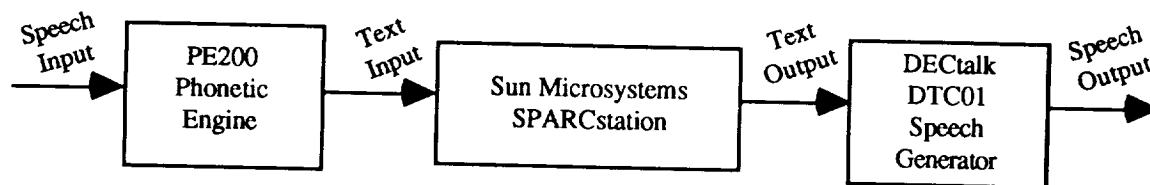


Figure 2
MAVT Architecture

DUAL-USE OF MAVT TECHNOLOGY

Other potential military applications of automatic spoken language translation include its use for multi-national military operations - facilitating communication among cooperative multi-lingual forces, and its use for deriving intelligence information from speech communications. This latter application requires a burdensome mix of knowledge-intensive skills similar to those involved in military interrogation since the analyst, or listener, must not only be able to listen to multiple lines of communication input and responsibly select and record information pertinent to the mission, but must also possess foreign-language skills when monitoring foreign-language communications.

The application of computer-based spoken language translation technology to language training can provide cost-effective augmentation and reinforcement of foreign language skills for inexperienced military linguists and civilian students. Students could request foreign language equivalents for spoken expressions in their native tongue or be encouraged to repeat phrases displayed on a screen or 'voiced' by the computer. The correctness of a student's attempt to speak in a foreign language would be determined by matching his utterance to the expected input. Instant reinforcement can be provided and intelligent computer prompting could guide students through difficult phrases. Adding other media, video for instance, to such a training system further extends the value of the technology as a training resource because students, cued by visual prompts, can then devise their own wording. Further, computer-based language training allows lessons to be varied depending on a student's level of competence.

The MAVT prototype has attracted the interest of law enforcement organizations and emergency room medical personnel. Most large metropolitan areas in the United States have many non-English speaking inhabitants, making communication and acquisition of information difficult in some cases and almost impossible in others. As a result of these communication problems law enforcement organizations such as the Los Angeles Police Department have expressed an interest in the use of automatic spoken language translation technology for interviewing crime witnesses, victims, and suspects. Hospital staffers have noted the value the technology holds for emergency room admittance of foreign-speaking patients. In a manner similar to the military application for interrogation, language translation technology would allow law enforcement personnel to communicate with citizens, and hospital personnel to communicate with patients, in their own languages without the time delay involved in locating an interpreter.

MAVT technology also finds application as a diplomatic, or business, briefing aid. Lack of a common language need not stand in the way, in the future, of international visits and communications among those with common political or business activities. Interest in the technology has been shown by a Texas organization eager to facilitate diplomatic interactions with representatives of the Mexican government.

Lastly, consider the value of the technology as a tourist travel aid. Currently, words or simple phrases can be typed into hand-held instruments that provide foreign-language translations, but how much better it would be to speak into a similar instrument and have it vocalize our intent in the language, or languages, of others.

TECHNOLOGY LIMITATIONS

The current state of the component technologies of the MAVT places limitations on the near-term application of computer-based spoken language translation. Some of the deficiencies of those technologies are identified here.

Speech recognition technology has improved substantially in just the last three years. Until very recently speaker independent continuous speech was still an out-of-range laboratory research goal, and the available isolated word speech recognition was not appropriate for very many applications. Current recognition capabilities will suffice, however, only for those applications for which every spoken input can be anticipated. Broader application of the technology will at least require the allowance of less restrictive, or more relaxed, grammars. In the meantime, the technology continues to advance rapidly.

Limitations set by the state of speech recognition technology on the breadth of the grammar actually make its use with text-based language translation viable since automatic translation technology is not yet capable of handling free-form text. Since every input is anticipated, correct translation of every possible input can fairly well be assured. Improvements in machine translation of text will be required in the future.

Broader application of speech translation technology will necessitate an improvement in intonation features offered by language generation systems. Computer-generated speech is currently robotic and monotone.

And finally, MAVT components are specifically tailored to operate with a specific language pair and within a specific domain. Tools to port spoken language translation technology to new domains and to new languages are needed.

REFERENCES

- Montgomery, Christine A, Bonnie G. Stalls, Robert E. Stumberger, Naicong Li, Robert S. Belvin, Alfredo Arnaiz, Philip C. Shinn, Armand G. DeCesare, and Robert D. Farmer. Machine-Aided Voice Translation (MAVT). To be published as a Rome Laboratory Final Technical Report.
- Montgomery, Christine A, Bonnie Glover Stalls, Robert E. Stumberger, Naicong Li, Sharon Walter, Robert S. Belvin, and Alfredo R. Arnaiz (May 24-27, 1993). Machine-Aided Voice Translation. In Proceedings of the 3rd Annual IEEE Mohawk Valley Section Dual-Use Technologies and Applications Conference, pp. 96-101.

2541
P. 8

**A PC PROGRAM TO OPTIMIZE SYSTEM CONFIGURATION FOR DESIRED RELIABILITY
AT MINIMUM COST**

Steven W. Hills
Idaho National Engineering Laboratory
Idaho Falls ID 83415-3765

Ali S. Siahpush
Idaho National Engineering Laboratory
Idaho Falls ID 83415-3765

ABSTRACT

High reliability is desired in all engineered systems. One way to improve system reliability is to use redundant components. When redundant components are used, the problem becomes one of allocating them to achieve the best reliability without exceeding other design constraints such as cost, weight, or volume. Systems with few components can be optimized by simply examining every possible combination but the number of combinations for most systems is prohibitive. A computerized iteration of the process is possible but anything short of a super computer requires too much time to be practical. Many researchers have derived mathematical formulations for calculating the optimum configuration directly. However, most of the derivations are based on continuous functions whereas the real system is composed of discrete entities. Therefore, these techniques are approximations of the true optimum solution. This paper describes a computer program that will determine the optimum configuration of a system for multiple redundancy of both standard and optional components. The algorithm is a pair-wise comparative progression technique which can derive the true optimum by calculating only a small fraction of the total number of combinations. A designer can quickly analyze a system with this program on a personal computer.

INTRODUCTION

Historically, systems have been designed and prototypes tested before the system reliability was analyzed. Some organizations use reliability engineers armed with component reliability data to analyze systems before prototyping, but the design engineer still does not have a tool to help in the preliminary design process. Customers of most modern systems expect a quantified reliability number and many engineers would like to design to specified reliability. In order for a designer to proceed toward a predetermined desired reliability, a tool for incorporating reliability analysis must be used early in the design process. The SYstem Reliability Optimizer (SYROP) is such a tool.

There are several methods commonly used to analyze system reliability. The one chosen for this application is the reliability block diagram (RBD). The RBD shows the success paths for the system which are generated from the functional block diagram. The connectivity of the blocks is often the same as the functional block diagram. Development of the functional block diagram is one of the first steps in a system design. Therefore, the RBD can also be produced very early in the design. The reliability of a component is defined as the probability of it performing the required function under stated conditions for a stated period of time. For components with constant failure rates this can be expressed as $R = e^{-\lambda t}$, where λ is the statistical failure rate and t is the time. With this type of data and the RBD, SYROP can be used very early in the design process to analyze the system reliability and optimize the configuration of the system. The SYROP model is quickly executed on a personal computer (PC) and is easily changed to accommodate the progress of the design or to perform parametric studies.

System reliability can be improved in several ways. Some procedures are implemented after the system is designed and fabricated. These include scheduled inspection, testing, and preventative maintenance. Two ways to improve reliability in the design phase are to use more reliable components and to use redundant components. When a designer uses one or both of these methods, the problem becomes one of optimizing the allocation of

redundant or improved components to achieve the necessary reliability without exceeding other design constraints such as cost, weight, or volume. For systems with few components and few options, this optimization can be simply performed by examining every possible combination. However, for most systems, the number of combinations is astronomical. For example, a system with 20 different components and 6 possible configurations for each component has 6^{20} , or 3.6×10^{15} possible combinations. That is why many researchers have derived mathematical formulations for calculating the optimum configuration directly. Some of the derivations have been based on continuous functions and others have attempted to account for the fact that component redundancy affects the system reliability in discrete increments. These approaches range from ones that simply rank the components in order of influence each one has on the overall system reliability to ones that attempt to optimize the redundancy allocation. Mohamed, Leemis, and Ravindran [1] provide a review of 62 papers and Tillman, Hwang, and Kuo [2] reviewed 144 papers describing techniques such as integer programming, mixed integer programming, dynamic programming, maximum principle, linear programming, geometric programming, sequential unconstrained minimization technique, modified sequential simplex pattern search, Lagrange multipliers, Kuhn-Tucker conditions, generalized Lagrangian function, generalized reduced gradient, heuristic approaches, parametric approaches, pseudo-Boolean programming, Hooke and Jeeves pattern search, and combinations of the above.

Designing for a specified reliability has also been discussed in the literature. Aggarwal [3] developed a method that would use the cost-reliability curve in the form of a mathematical function. Aggarwal and Sharma [4] proposed that the PC could be used with an incremental technique. Rao and Dhingra [5] developed two optimization techniques coupled with heuristic approaches to solve the mixed integer nonlinear programming problem of multistage systems. Numerous other papers can be found on the subject. The problem is that none of these approximations has produced a practical tool for the designer working with actual components that have distinct physical properties such as cost, weight, volume, and reliability.

A computer program is the obvious answer to analyze all the combinations and determine the true optimum solution. However, calculation of the system reliability and cost for the many (e.g. 3.6×10^{15}) combinations would take a PC several hours and a super computer is not readily available to most engineers. Therefore, an effort was made to develop an algorithm or procedure that would determine the true optimum redundancy allocation while minimizing at least one system constraint without having to calculate every combination. This program must be based on using actual component reliability and physical parameters available from vendor data or other engineering analyses.

PAIR-WISE COMPARATIVE, PROGRESSIVE DOMINANCE ALGORITHM

For this analysis, all component failures are assumed to be independent and redundant components are assumed to be in parallel, not standby mode. The reliability of the block (in the RBD) to which a redundant component is added will be increased by the second component functioning in parallel because failure of that block would require both components to fail. If the reliability of a component is R , the reliability of two identical components in parallel is $R' = 1 - (1 - R)^2$ and the reliability of three identical components in parallel is $R'' = 1 - (1 - R)^3$.

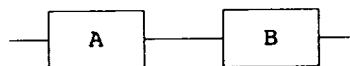


Figure 1. A Two-Component System.

Consider a system of two serial components whose RBD is shown in Figure 1. System failure would occur if either one failed because the success path through the RBD must include both A and B. The reliability of the serial system is the product of the reliability of each component; $R_s = R_A R_B$. In other words, the reliability of the two components in series is less than the reliability of either one alone. In order to increase the system reliability redundant components may be added to either A or B, components of higher reliability may be substituted for either A or B, or a combination of redundant and optional components may be used. SYROP is arbitrarily limited to adding no more than two redundant components to each block. Two redundant components

implies that there are three identical components in parallel. This limit was imposed because more than three parallel components will not add significant reliability to a system unless the component reliability is very low. The number of optional components is limited to one. Therefore, with both standard and optional components available and with redundancies of 0, 1, or 2 possible for each, there are 6 potential configurations for each block of the RBD. Six configurations for each block results in 36 possible configurations for the two-component system.

Next, examine the system in Figure 1 with some assumed values of component reliability and cost. Assume that there are standard and optional components available for both A and B with values as listed at the top of Figure 2. The 36 possible combinations for this system are also shown in Figure 2 and the system reliability versus cost is plotted in Figure 3 with each point labeled by its configuration number. Investigation of this graph reveals that the sequence of 10 configurations connected by the dashed line "dominates" all other configurations. One configuration dominates another if it has higher reliability at no more cost or no less reliability at less cost. For instance, configurations 4, 10, and 19 have equal cost but 4 has higher reliability and is, therefore, the dominant one. Also, configurations 3, 16, and 22 have equal cost but configuration 5, which has lower cost, dominates because it has higher reliability. Kettelle [6] showed that, for a system of serial components, the complete set of dominant configurations is composed of subsets of the dominant sequences of subsystems of the whole system. It can also be shown that this holds true for all subsystems, whether they are in series or parallel, as long as each subsystem is either all parallel or all serial [7].

The pair-wise comparative, progressive dominance algorithm used in SYROP calculates reliability and cost for only two components or nodes at a time (hence, the term "pair-wise comparative") and progresses from the lowest level of the model to the top level of the model, selecting the dominant sequences at each node (hence, the term "progressive dominance"). Only the dominant sequence of configurations is used to calculate the reliability and cost of the next higher level. This allows the dominant sequence of the top level to be determined from a much smaller number of calculations than if all configurations were calculated. As systems increase in size, the required number of calculations may increase but the ratio of required number to total possible will decrease. With this algorithm a PC can analyze a system in seconds that would take hours if all combinations were calculated.

COMPUTER PROGRAM

The nodal model is developed by first dividing the system into serial "stages" where each stage is the smallest group of components that can be kept in series with other stages. Each stage is then broken into parallel "legs" of components and each leg is broken into serial "subsystems" and each subsystem is separated into parallel components. This will be illustrated by the following hypothetical example. Suppose that a system consists of twelve components and its RBD is drawn in Figure 4. The designer has made an initial selection of standard components with reliability and cost data as listed in Table 1. Seven of the components also have optional values to consider. This RBD has three serial stages. The first stage is only component 1, the second stage contains components 2 through 6, and the third stage is components 7 through 12. Each of these stages is the minimum set of components that can be grouped in series with each other. The first stage has no subsets, but the second stage has two parallel legs of subsystems and the third stage has three parallel legs of subsystems. The subsystems of all legs are single components except for the bottom leg of stage 2. The second subsystem in this leg is composed of two components in parallel.

It takes just a few minutes to input this data to SYROP and then it will determine the sequence of configurations that dominate all possible combinations. The combinations include six different configurations for each of the first seven components but only three configurations for components 8 through 12 because there are no optional devices available. Therefore, the total number of combinations is $6^7 \cdot 3^5 = 68,024,448$. Of course, the pair-wise comparative, progressive dominance algorithm does not make that many calculations and it takes an 80386 PC about four seconds to determine the dominant sequence and write the data files for this example. The first output data is a plot of the 82 dominant configurations as shown in Figure 5. The dominant sequence ranges from the initial configuration with no redundant or optional components ($R_s = .8068$, $C_s = 13$) to the one with

		<u>Reliability</u>	<u>Cost</u>
Component A	Standard	0.80	100
	Optional	0.85	200
Component B	Standard	0.75	150
	Optional	0.85	250

36 Possible Configurations: S denotes a standard component and O denotes an optional component.

<u>Configuration</u>	<u>Reliability</u>	<u>Cost</u>	<u>Configuration</u>	<u>Reliability</u>	<u>Cost</u>
(1)	.600	250	(10)	.680	350
(2)	.750	400	(11)	.782	600
(3)	.787	550	(12)	.797	850
(4)	.720	350	(13)	.816	450
(5)	.900	500	(14)	.938	700
(6)	.945	650	(15)	.957	950
(7)	.744	450	(16)	.843	550
(8)	.930	600	(17)	.970	800
(9)	.977	750	(18)	.989	1050

Figure 2. Possible System Configurations for the Two-Component Example

Component A	Standard	<u>Reliability</u> 0.80	<u>Cost</u> 100
	Optional	0.85	200
Component B	Standard	0.75	150
	Optional	0.85	250

36 Possible Configurations: S denotes a standard component and O denotes an optional component.

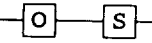
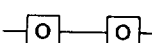
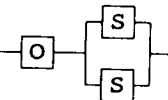
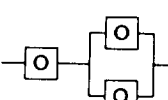
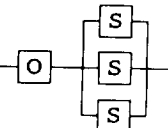
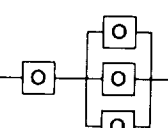
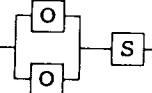
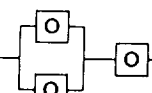
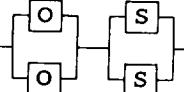
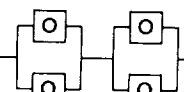
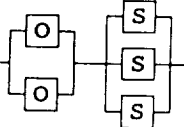
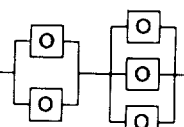
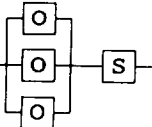
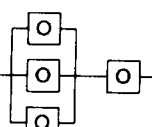
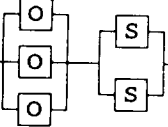
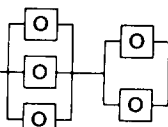
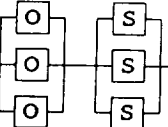
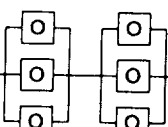
<u>Configuration</u>	<u>Reliability</u>	<u>Cost</u>	<u>Configuration</u>	<u>Reliability</u>	<u>Cost</u>
(19) 	.638	350	(28) 	.723	450
(20) 	.797	500	(29) 	.831	700
(21) 	.837	650	(30) 	.847	950
(22) 	.733	550	(31) 	.831	650
(23) 	.916	700	(32) 	.956	900
(24) 	.962	850	(33) 	.974	1150
(25) 	.747	750	(34) 	.847	850
(26) 	.934	900	(35) 	.974	1100
(27) 	.981	1050	(36) 	.989	1350

Figure 2. (continued) Possible System Configurations for the Two-Component Example

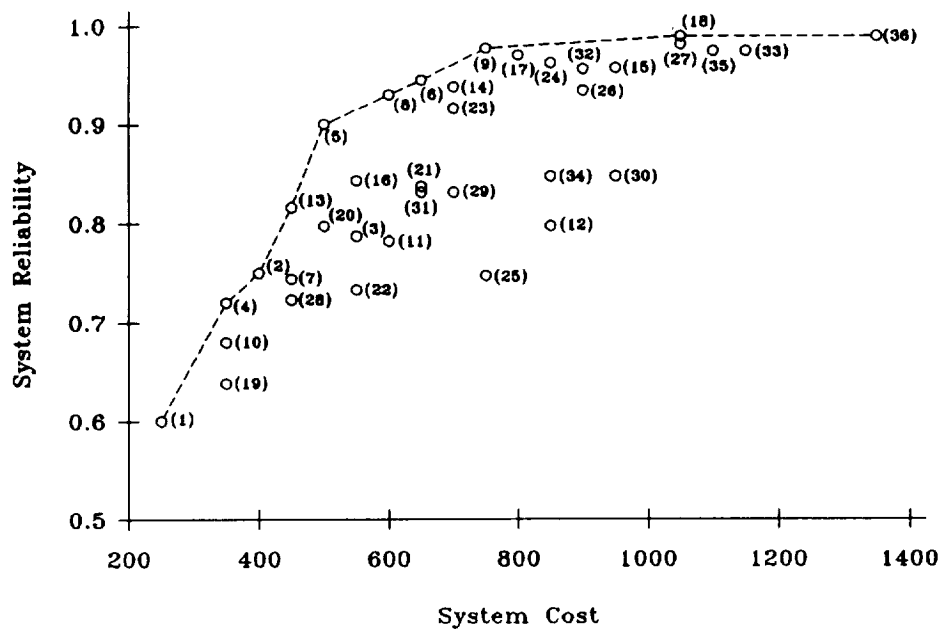


Figure 3. Reliability vs. Cost for the Two-Component Example.

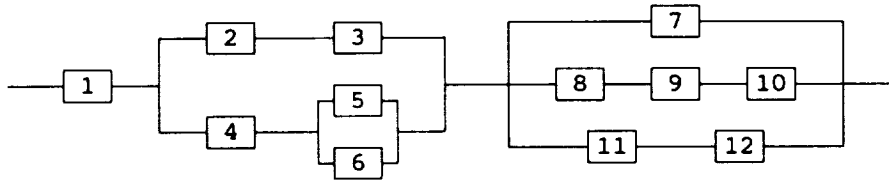


Figure 4. Reliability Block Diagram for the Hypothetical Example.

Component Number	Standard Values		Optional Values	
	Reliability	Cost	Reliability	Cost
1	.90	1.0	.95	2.0
2	.80	1.0	.90	2.5
3	.80	1.0	.95	3.0
4	.80	1.0	.85	1.5
5	.75	1.5	.80	2.0
6	.75	1.5	.80	2.0
7	.85	1.0	.95	2.0
8	.85	1.0	-	-
9	.90	1.0	-	-
10	.90	1.0	-	-
11	.80	1.0	-	-
12	.85	1.0	-	-

Table 1. Reliability and Cost Data for the Hypothetical Example

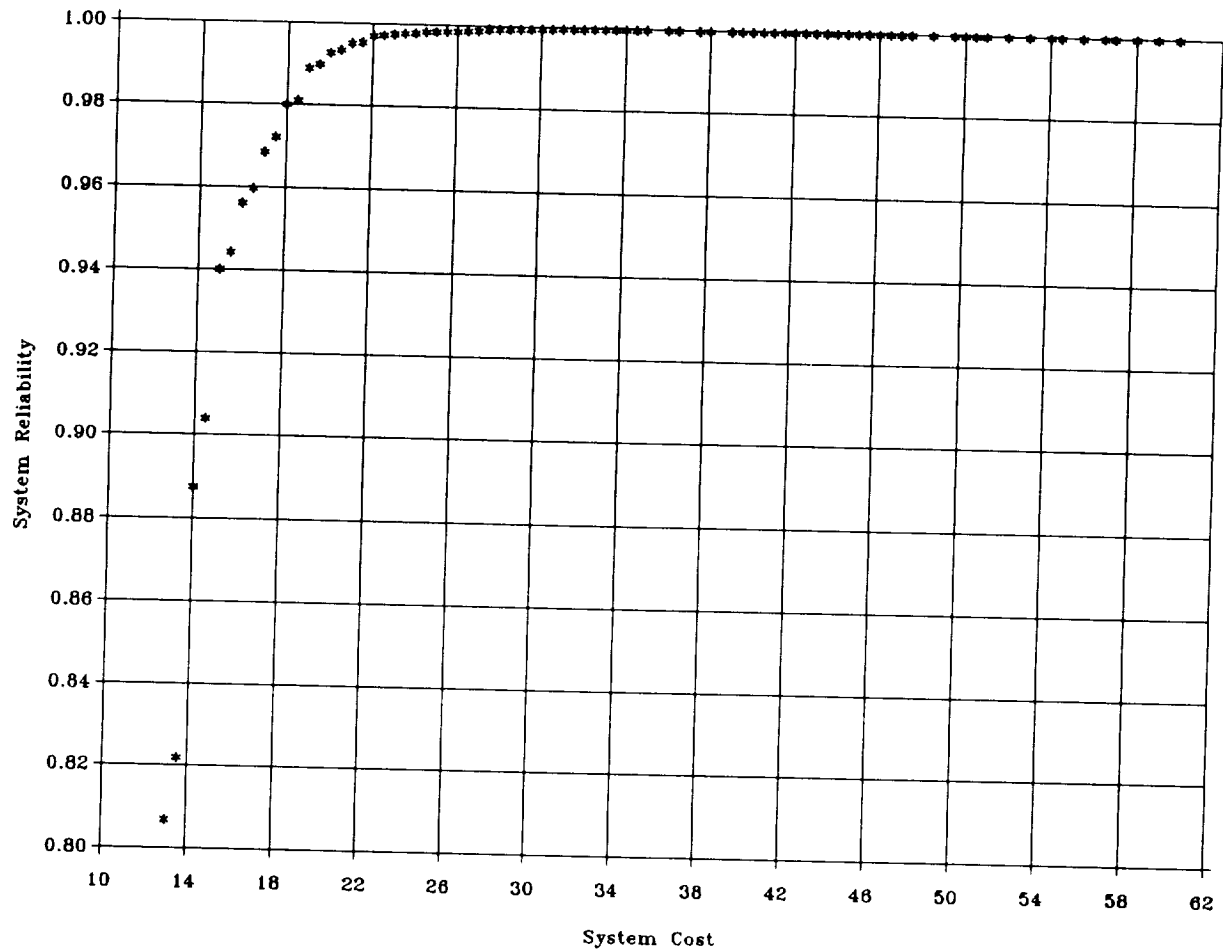


Figure 5. Reliability vs. Cost for the Dominant Configurations of the Hypothetical Example.

double redundancy for all components and optional ones where available ($R_s = .9999$, $C_s = 60$). This picture very quickly illustrates the reliability/cost options for the system. SYROP then displays the system configuration details when the user inputs the desired system reliability. For example, if a desired reliability of 0.97 is selected, SYROP displays the configuration as shown in Figure 6. The calculated reliability of the system for this configuration is 0.972 and the cost is 17.5. This configuration includes redundant standard components for components 1, 2, 3, and 7 and a single optional component in place of number 4. The other components are unchanged. Additional values of desired reliability can be quickly investigated once the dominant sequence is determined. The user can also print a list of the dominant sequence from lowest reliability to highest reliability with the configuration of each component given. This allows the system designer to see any trends that may be in the system. The data in SYROP can be easily changed to make a parametric study of the system components.

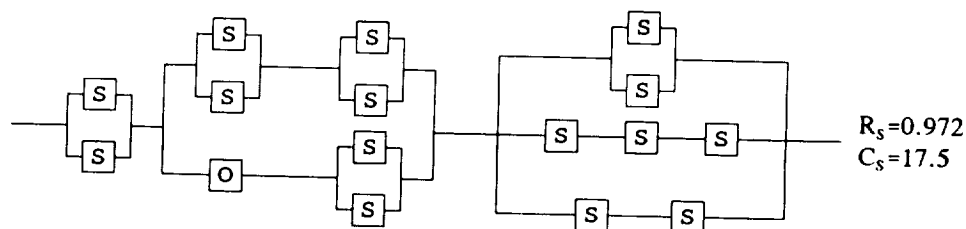


Figure 6. System Configuration of Hypothetical Example for Desired Reliability of 0.97.

FURTHER WORK

There are several things planned to improve SYROP. The most basic is the user interface. The data input graphics, file management, and information output are currently being enhanced. Other functional improvements that are under consideration include:

Limit redundancy for specific components. There may be physical constraints on some components of a system that do not allow redundant components or may allow only one redundant component. SYROP could have the option to set the redundancy limit for specific components.

Link two blocks of the RBD together logically. The functional block diagram of a system may dictate that the same block appears twice in the RBD. If this is the case, the configuration of both of these blocks would have to be the same in order to be physically correct.

Consider mixed redundancy. For some systems there may be an advantage to have a redundant component that is different than the initial component rather than forcing all redundant components to be the same, as SYROP currently does. This could be especially true for standby components.

Analyze standby systems. All calculations in SYROP currently assume pure parallel redundancy. An analysis technique for standby systems would be useful.

Analyze dual mode failure. Some components may have two modes of failure. For example, hydraulic valves or electrical switches may fail to close on demand or fail to open on demand and the time dependent or demand dependent failure rate is often different for each mode. A technique has been developed [7] to analyze dual mode failure and it may be possible to incorporate this into SYROP.

In addition to the ideas mentioned above, SYROP needs to be evaluated for several actual systems to determine its strengths, weaknesses, and applicability. Development will continue in order to establish this as a useful tool for design engineers.

REFERENCES

1. Mohamed, A.A., Leemis, L.M., and Ravindran, A., "Optimization Techniques for System Reliability: A Review", *Reliability Engineering and System Safety*, **35** (1992) 137-146.
2. Tillman, F.A., Hwang, C.L., and Kuo, W., "Optimization Techniques for System Reliability with Redundancy - A Review", *IEEE Transactions on Reliability*, **R-26** (3) (1977) 148-155.
3. Aggarwal, K.K., "Minimum Cost Systems with Specified Reliability", *IEEE Transactions on Reliability*, **R-26** (3) (1977) 166-167.
4. Aggarwal, K.K. and Sharma, S., "Microprocessor Based Redundancy Designer", *Reliability Engineering and System Safety*, **31** (1991) 391-398.
5. Rao, S.S. and Dhingra, A.K., "Reliability and Redundancy Apportionment Using Crisp and Fuzzy Multiobjective Optimization Approaches", *Reliability Engineering and System Safety*, **37** (1992) 253-261.
6. Kettelle, J.D., "Least-Cost Allocations of Reliability Investment", *Operations Research*, **10** (2) (1962) 249-265.
7. Hills, S.W. and Siahpush, A.S., "SYROP: A Tool for Incorporating System Reliability in the Design Phase", Technical Report EGG-ME-11052 (in preparation).

**Evolving Software Reengineering Technology
For the Emerging Innovative-Competitive Era**

2502
P-10

**Phillip Q. Hwang
Defense Mapping Agency
8613 Lee Highway
Fairfax, VA 22031-2137**

**Evan Lock, Noah Prywes
Computer Command and Control Company
2300 Chestnut Street, Ste. 230
Philadelphia, PA 19103**

ABSTRACT

This paper reports on a multi-tool commercial/military environment combining software Domain Analysis techniques with Reusable Software and Reengineering of Legacy Software. It is based on the development of a military version for the Department of Defense (DOD). The integrated tools in the military version are: Software Specification Assistant (SSA) and Software Reengineering Environment (SRE), developed by Computer Command and Control Company (CCCC) for Naval Surface Warfare Center (NSWC) and Joint Logistics Commanders (JLC), and the Advanced Research Project Agency (ARPA) STARS Software Engineering Environment (SEE) developed by Boeing for NAVAIR PMA 205. The paper describes transitioning these integrated tools to commercial use. There is a critical need for the transition for the following reasons: First, to date, 70% of programmers' time is applied to software maintenance. The work of these users has not been facilitated by existing tools. The addition of Software Reengineering will also facilitate software maintenance and upgrading. In fact, the integrated tools will support the entire software life cycle. Second, the integrated tools are essential to *Business Process Reengineering*, which seeks radical process innovations to achieve breakthrough results. Done well, process reengineering delivers extraordinary gains in process speed, productivity and profitability. Most important, it discovers new opportunities for products and services in collaboration with other organizations. Legacy computer software must be changed rapidly to support innovative business processes. The integrated tools will provide commercial organizations important competitive advantages. This, in turn, will increase employment by creating new business opportunities. Third, the integrated system will produce much higher quality software than use of the tools separately. The reason for this is that producing or upgrading software requires keen understanding of extremely complex applications, which is facilitated by the integrated tools. The radical savings in the time and cost associated with software, due to use of CASE tools that support combined Reuse of Software and Reengineering of Legacy Code, will add an important impetus to improving the automation of enterprises. This will be reflected in continuing operations, as well as in innovating new business processes. The proposed multi-tool software development is based on state of the art technology, which will be further advanced through the use of open systems for adding new tools and experience in their use.

1. Introduction And Summary

The paper describes a multi-tool Software Engineering Environment for commercial and military applications combining software Domain Analysis techniques with those of Reusable Software and Reengineering of Legacy Software.

The paper is based on the dual use of a version developed for the Department of Defense (DOD). The tools used in the military version must be modified for use in a commercial version. They are as follows: Software Specification Assistant (SSA) and Software Reengineering Environment (SRE) developed by Computer Command and Control Company (CCCC) under contracts with the Naval Surface Warfare Center (NSWC) and Joint Logistics Commanders (JLC), and the ARPA STARS Software Engineering Environment (SEE) developed by Boeing for NAVAIR PMA 205.

The *Software Reengineering Environment* (SRE) will facilitate:

- a) Translation of legacy software from old languages (Fortran, Cobol, C) to modern languages (C++ and Ada) and a modern open-ended Operating System (OSF).
- b) Software Understanding and Documentation through analysis of relations within the software and their graphical display.
- c) Reorganization of software to obtain object oriented programs for concurrent network operations.

The *Software Engineering Environment* (SEE) will facilitate:

- a) New software design.
- b) Reuse of code associated with a Domain.
- c) Code generation.

This system is much more important than just combining capabilities of tools. It is a change in Computer-Aided Software Engineering (CASE) technology that will radically improve overall business automation. It will drastically lower the cost and time to develop software, for the following reasons:

First, 70% of programmer time is applied to software maintenance [5]. The work of these users is not facilitated by existing CASE tools. The Software Reengineering tools proposed here will also facilitate the work of users engaged in software maintenance and upgrading. The integrated tools will have a much wider coverage of the software life cycle and a much wider audience.

Second, the integrated Software Engineering and Reengineering are essential to *business process reengineering*, which seeks radical process innovations to achieve breakthrough results [25], [42]. Done well, process reengineering delivers extraordinary gains in speed, productivity and profitability. More importantly, it discovers new opportunities for products and services offered in collaboration with other organizations. Rapid offering of the new products and services is essential, in order not to miss business opportunities. Innovation in legacy computer software has been recognized in studies of reengineering as the major obstacle in rapidly reengineering manufacturing processes. Software must change rapidly to support the core requirements of business process engineering. These requirements are [27], [24]:

- a) Support of concurrent operations
- b) Leveraging human resources
- c) Sharing information
- d) Collaborating with other organizations

The transitioned tools will offer commercial organizations the necessary means to attain strong, competitive advantages. This, in turn, will impact employment favorably by creating new business opportunities.

Third, the integrated tools offer a new, higher quality of software than the tools offer separately. The reason for this is that producing or upgrading software requires keen user understanding of extremely complex applications, which is facilitated by the Software Reengineering tool [35].

The transitioning from a military to commercial environment requires the following changes (only components of the military version that require changes for the commercial version are listed here):

- i) Translation of commercial programming languages Fortran, Cobol, C, and C++ into C++. The present translations in the military version are from CMS-2 and Ada to Ada. C++ is selected as the modern programming language preferred by the commercial community.
- ii) Integration of CCCC's SRE tools with those of Domain Engineering and Application (DEA) [40], developed by Boeing under ARPA/STARS sponsorship, and with that of the PTECH tool [38] for Object Oriented Software De-

sign. The integrated tools will be transitioned to operate under the Open Software Foundation (OSF) Operating System. This will make them portable to multiple vendors' platforms. The military version operates on Digital's Vaxstations under the VMS Operating System.

2. Overview Of The System

Figure 1 on the following page illustrates the technical approach. It shows the integrated operation of the Software Reengineering Environment (SRE) with Domain Engineering and Application (DEA) and PTECH, an object oriented CASE tool. The integrated operation covers the software life cycle for commercial use. Figure 1 shows the integration of three main tool groups, as follows:

Software Specification Assistant (SSA): It is shown at the top left of Figure 1. SSA facilitates the creation and updating of software requirements and specifications. In the military version it conforms with DOD-STD 2167A [19]. A similar set of standards will be selected for the commercial version. The inclusion of SSA reflects the importance of Software Specifications for an orderly software life cycle. SSA is an integrated set of information repositories and tools. SSA guides, instructs and informs staff in composing, updating and evaluating preliminary requirements and specifications. Typical users of SSA are a Development Manager, Software Support Activities, or Contractors. SSA allows a user to manage, query and update its extensive knowledgebase of information related to an application system. Staff may ask complex technical questions about the software and retrieve answers. Fragments of retrieved answers can be extracted for inclusion in updates to relevant new documents. SSA leads the inexperienced specifier in a "step-by-step" manner and provides traceability to the sources for updates in the documentation. This component is very flexible and open ended and does not require changes (beyond defining a commercial standard) to transition it to the commercial version.

SSA has been used in a demonstration project for the Tactical Air Mission Planning System (TAMPS) program at the Naval Air Warfare Center, Warminster, PA. It is planned to distribute the SSA tool to DOD agencies.

Software Engineering Environment (SEE) [23]: It is shown at the top right of Figure 1. SEE incorporates new software development technology for Reuse of software and for automatic program generation, following ARPA's Domain Specific Software Architecture (DSSA) Program [31]. The ARPA STARS SEE for NAVAIR PMA 205, used in the military version, includes two parts: i) Domain Engineering, to define the process of producing software for a class of related applications in a Domain, and ii) Application Engineering, for producing software for an application that belongs to the Domain [40]. It also contains PTECH [38] for object oriented design of software. The SEE can be transitioned to the commercial version without any changes. DEA is language independent and PTECH already generates object oriented programs in C++. The SEE incorporates meta tools for tool integration.

A specific Domain is comprised of software for a closely related family of applications. For example, the DSSA program has been experimenting with Domains in the areas of Guidance, Navigation and Control, Avionics, Command Control and Communications, etc. [8] Once a Domain Software Architecture [1] is developed, it can be used for generating applications. Business Domain architectures will be developed by respective users for their areas of interest, and will be populated by artifacts from legacy code. Domain Engineering contains a repository of Reuse Software artifacts and associated tools. The creator of a Domain Architecture is called the *Domain Engineer*. The DEA facilitates selection of Reuse Software and generation of software to create a specific application system. The user of the repository and of the tools is called the *Application Engineer* [40].

The Reuse Software is part of the DEA Repository (see Figure 1). It is organized as a hierarchy of software artifacts that follow the Domain Architecture. The DOD software also follows a standard structure of hierarchical software units (DOD-Standard 2167A). Each hierarchical software unit has a specification of its position in the architecture hierarchy, its capabilities, interfaces and dependencies on other software units. The graphs for documentation of the architecture are listed in Table 1. The capabilities of the hierarchical software units determine *commonality* and *variability* between hierarchical software units. It is possible to navigate through the Domain hierarchy tree by referring to capabilities and selecting hierarchical software units based on commonality and variability of their respective capabilities. Hierarchical software units may be parametrized and a code generation tool may be used to select parameters of generic software. Alternately, hierarchical units may be completely generated based on models of their functionality.

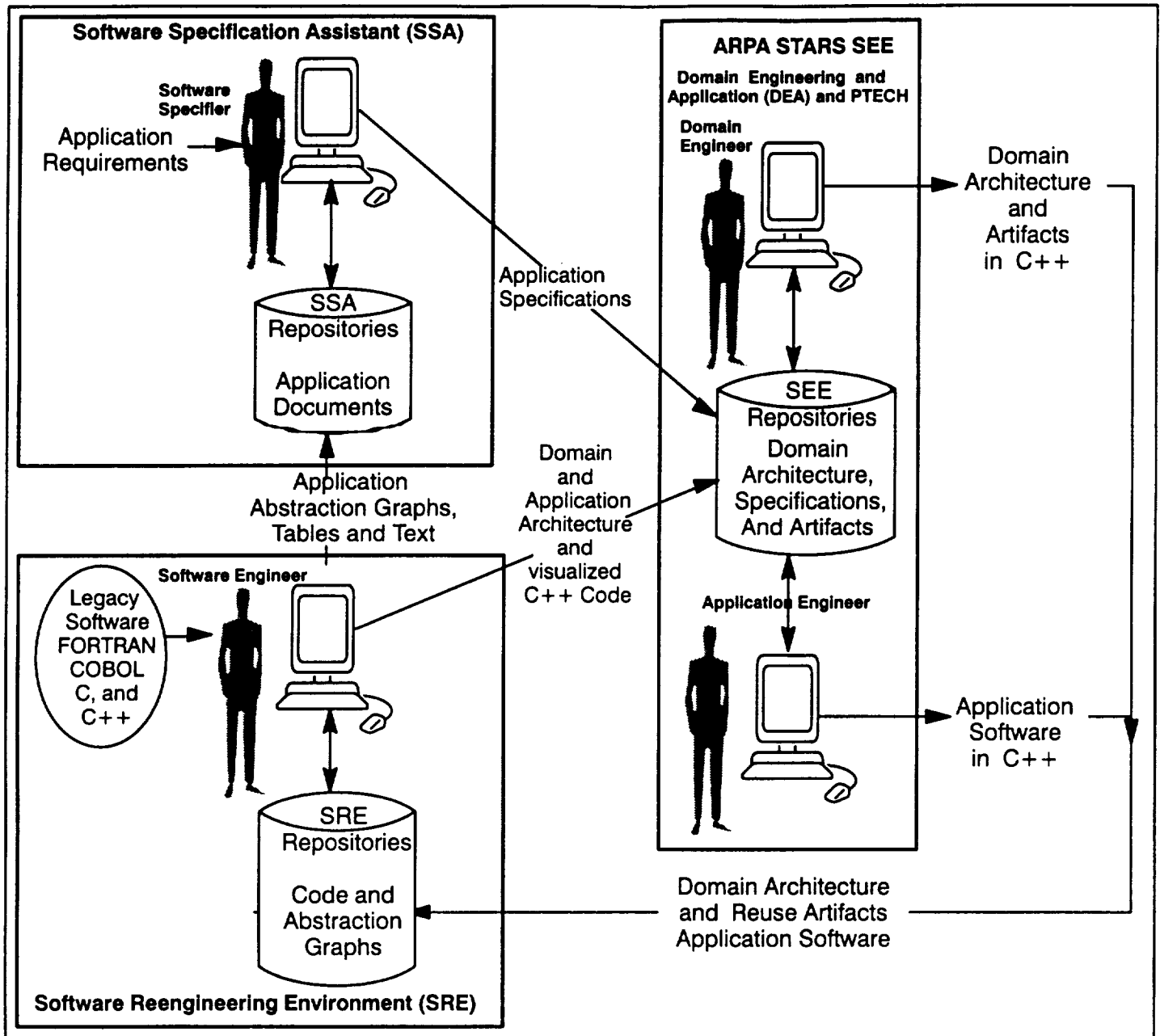


Figure 1: Overview of Integrating Tools for the Software Life Cycle.

A series of tools are available in the DEA for application modelling, unit testing and conversion to concurrent operation:

Software Reengineering Environment (SRE) [9]: SRE is shown at the bottom left of Figure 1. The SRE commercial version will consist of three main capabilities:

- (i) *Software Understanding*: It consists of query and retrieval of graphic diagrams that illustrate the software from various perspectives. These diagrams are used to visualize specific aspects of the software. The diagrams are first divided into *in-the-large* and *in-the-small* diagrams. In-the-large diagrams visualize declarations of objects. In-the-small diagrams visualize execution statements within individual program units. The C++ program diagrams will be stored in a graphic form in the repository of a customized CASE system. A graphic query language is provided for ad-hoc browsing in the graphic repository, consisting of the documents in Table 1. These graphs show relations between high or low level hierarchical units. This facilitates the understanding of the software's architec-

ture as well as its code. Changes to the program can be made via the graphics used for visualization. This part of the military version can be used directly with no changes for the commercial version. The changes to C++ are all in the translation component (i).

- (ii) *Software Abstraction and Documentation*: SRE partitions the software into multi-level hierarchical software units, in conformity with the standards for describing the software architecture. Software documents are then generated that describe the architecture in terms of the hierarchical software units, and describe the software in each unit from different perspectives shown in Table 1 below. This part of the military version can be used directly in the commercial version.
- (iii) *Software Capture and its Transformation to C++ and OSF*: Fortran, Cobol, C or C++ source programs will be translated into C++ Entity-Relation-Attribute (ERA) diagrams. This representation is the main vehicle for graphic program analysis and visualization. Multiple passes are made over the source code to achieve 100% translation to object-oriented C++. Visualization is used for query, retrieval, understanding and generating documentation of programs. In a first pass, the SRE will translate the code, statement by statement, into pseudo-C++. Next, the generated pseudo-C++ programs will be transformed into the C++ programming paradigm in a series of passes. Each pass translates different aspects of the programming paradigm of the source language into the C++ programming paradigm (e.g., object declarations and classes). The Unix commands in legacy code will be replaced by OSF code having the same functionality. Other commercial subsystems (user interface and database) will be translated in this way as well. The military version uses Ada as the target language. It will have to add C++ as the target language.

(i)	Hierarchical Decomposition Diagrams	Showing decomposition of the overall software into hierarchical units.
(ii)	Flow Diagrams	Showing flow of data and control within and between hierarchical units.
(iii)	Interface Tables	Showing the structure of inputs and outputs of each hierarchical unit.
(iv)	Object/Use Diagrams	Showing (for hierarchical units) where types or generics are defined and where they are used.
(v)	Context Diagrams	Showing the library units and where they are used.
(vi)	Comments Text	Showing the comments in each hierarchical unit. They are assumed to contain information on the hierarchical unit's capabilities.

Table 1: Software Abstraction Documents Produced by SRE for Different Perspectives of the Software.

The integration of the SSA, SRE, DEA and PTECH involves primarily two interfaces, also shown in Figure 1. They are as follows:

Interface Between SRE and SSA: This interface is shown at the middle left of Figure 1 [37]. This interface provides a *reverse* process to produce information for the software requirements/specifications and other documentation from program code. SSA receives the documentation from SRE.

Interface Between SRE and SEE: The SRE provides DEA and PTECH with software documentation in the form of high level graphic-views of the architecture as well as detailed graphic views of algorithms. The SRE can process Legacy code as well as Reuse code from the DEA repository. It generates key parts of the specifications of each hierarchical software unit. The capabilities of each hierarchical software unit in the specifications are used for establishing commonality and variability among the domain architecture hierarchical software units.

Discussion of the Software Life Cycle Process using the integrated tools: The visualization provided by SRE facilitates human understanding. This is essential for successful employment of the entire Domain Specific Software Architecture (DSSA) concept. These SRE capabilities are needed for:

- i) Analyzing and understanding the programs in the Reuse software library of Domain Engineering and in the target software produced by Domain Application.
- ii) Analyzing, transforming and understanding existing software to expand the domain architecture with new artifacts. This capability includes the transformation of existing programs in other source languages to C++.

DEA will use the SRE visualization graphs to compare the domain architecture's unit and assess commonality and variability of hierarchical software units. The SSA will be used to specify the capabilities of the architecture's hierarchical software units.

Typically the tools will be used iteratively until a desired new or upgraded application system is obtained. As an example, consider the following scenario. Assume for simplicity that totally new application software is desired. The preliminary requirements have been composed by the application's Program Manager. The platform to be used and its dynamics may be derived through modelling and simulation. The Specifier, with the aid of SSA, uses the preliminary requirements to compose the hierarchically structured specifications. The capabilities in the specifications are then communicated to the Application Engineer to select architecture units and generate the application software. If unable to do so, the Domain Engineer may be called to expand the scope of the domain. In either case, the SRE tool will be used to document and display the new domain and/or application software. Software Abstractions will be generated from the code. The Software Abstractions are next used by the Specifier, who employs SSA to update the specifications. The Domain Engineer will use the Software Abstractions to update the domain. The Application Engineer will use them to document the application software. This cycle may be repeated a number of times until satisfactory application software is realized.

3. Discussion Of The Transition

A military version of the system is partly operating (SRE) and partly in development (DEA). The transition into a commercial version will use the following technologies (The components of the military version which do not require alteration for the commercial version are not included):

- i) Translation from older commercial languages (Fortran, Cobol, C) and from C++ to C++ and from Unix to OSF (instead of Ada in the military version).
- ii) Use of meta-languages for assembling multiple tools. They must be transitioned from the Digital VAX workstations and VMS Operating System used in the military version, to the OSF Operating System, which is portable among different vendors' workstations.

3.1 Translation Technology

The methodology used in the present military version of SRE for translation of military source languages to Ada will be transitioned to translate Fortran, Cobol, C, and C++ into C++ and Unix to OSF. The graphic representation of programs in C++ is similar to the graphic representation of Ada. DEA is language independent of the target language and PTECH produces C++ code as well as Ada.

The processing in the SRE is shown in Figure 2 on the following page. SRE consists of four phases: *Parsing and Transformation (P&T)*, *Analysis and Restructuring (A&R)*, *Software Understanding (SU)* and *Software Abstraction (SA)*. Only A&R will need to be revised.

The data is stored in three repositories: Software Reengineering (SR), Software Understanding (SU) and Software Abstraction (SA).

The input to P&T is a source language program. The output of P&T will be in Elementary Statement Language for C++ (ESL-C++). This is the graphic notation for visualizing C++ programs. An approach very similar to the one used in Ada is envisaged. Namely, the relations represented by edges in Ada can be retained in C++. These are: i) Edges between caller of a procedure and the procedure. ii) Edges between memory updating or referencing a variable and the variables' declaration. iii) Edges between message source and its destination. iv) Edges between a class and its instances. It is stored as a tree in the SR Repository in ASCII form, and may be modified by the users updating the programs.

A&R modifies and enhances the ESL-C++ tree in the SR Repository. It adds *tuples* that represent relations between statements. The results are restored in the SR repository. Visualization views are generated by A&R and stored in the SU Repository that is used for display by a CASE system.

The DECdesign CASE system is being used for the SU tool. The Graphic User Interface (GUI) used in SU is customized for software reengineering and understanding. SU supports graphic retrieval and generation of C++ code. Once the programs are generated in C++, they can be added to a DEA Reuse Repository.

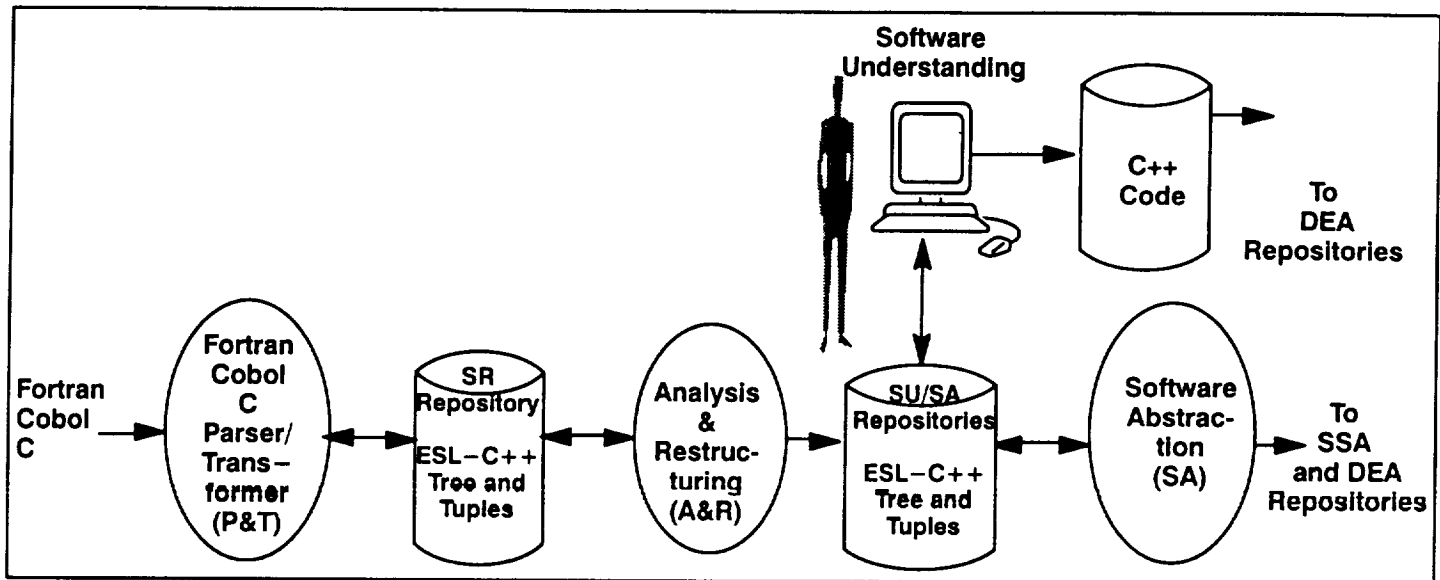


Figure 2: Tools and Repositories in the SRE.

SA produces diagrams, tables and text for the different perspectives of the software. These can be inserted in specifications of respective hierarchical software units. The SA Repository views are generated from the program visualization stored in the SU Repository. The SA Repository also uses formats acceptable by the CASE system used in SU.

3.2 Integrating Tools and Sharing Repositories

This section discusses systems used for Integration of Tools and Repositories. Three technologies are used in the ARPA STARS SEE. These systems are currently operational on Digital's Vaxstations under VMS. They need to be ported to OSF. They are as follows:

- i) Process driven technology to combine operations of tools. Honeywell's AAA system is used [29].
- ii) Tool and Repository integration technology. Digital's A Tool Integration Service (ATIS) [14] and HP-Softbench are used.
- iii) Repository data dictionary technology for sharing data in a repository. Presently Digital's Common Data Dictionary/Repository (CDD/R) is used under VMS. It will be changed to use a vendor independent repository under OSF (Digital's Objectivity is a primary choice).

To incorporate new tools and repositories into the SEE, the schemas in ATIS or HP-Softbench and the Data Dictionary must be extended to include the following:

- i) Declarations of new element types of the repositories that are managed by the tools.
- ii) Declarations of the tools.
- iii) Declarations of messages and methods that interface between user programs, the tools and the repository objects.

Finally, the processes that combine the use of multiple tools and a common user graphic interface must be redefined.

4. Conclusion

The key requirement in the emerging era will be to perform major software reengineering over a few days, at most. This will require extending the man-machine interaction of the human participants in the process. The approach described in Figure 1 can be expanded by adding support for multiple source and target programming languages. The proposed requirements are shown in Table 2. They consist of adding Cobol, Fortran, C and C++ as source program languages for the commercial world, and C++ as a target program language.

SRE: Software Reengineering	Accept Multiple Source Languages:	Military (CMS-2, Jovial, Ada) Commercial (Cobol, Fortran, C, C++)
	Generate Multiple Target Languages:	Ada, C++
	Reorganization:	For Object Orientation
	Understanding:	Graphical Query and Retrieval
	Abstraction:	Follows User Partitioning of Software Multiple Diagrams of Software Architecture: Decomposition, Flow, Objects and their instantiations
SSA: Software Specification	Very Large Repository of Documents (Text and Figures),	
	Very Rapid Search, Edit and Compose Facilities	
	Document Loading	
	Process Management	
	Portability Option	
SEE: Software Reuse and Generation	Domain Engineering Driven From Specifications	
	Domain Application Driven From Commonality/Variability	
	Structured Repository of Code and Specification	
	Code Generation	
	Concurrency Analysis	
	Simulation	

Table 2: Specification for Software Reengineering/Development For Emerging Era.

5. References

1. Agrawala, A., et al, "Domain-Specific Software Architectures for Intelligent Guidance, Navigation & Control," Proceedings of the DARPA Software Technology Conference 1992, Los Angeles, CA, April 1992.
2. Biggerstaff, T. J. "Design Recovery for Maintenance and Reuse," *IEEE Computer*, July 1989, pp. 36-49.
3. Biggerstaff, T. J., ed. *Software Reuse*, Addison Wesley 1987.
4. Black, E., "White Paper: ATIS, PCTE, CIS and Software Back Plane," Atherton Technology, Sunnyvale, CA 94089, 1991.
5. Bloom, P., "Case Market Analysis," Volpe, Welty and Co., San Francisco, CA 1990.
6. Chang, S., *Visual Languages and Visual Programming*, Plenum Press, 1990.
7. Chen, P., "The Entity-Relationship Model: Toward A Unified View of Data," *ACM Trans. on Database Systems*, May 1976.
8. Coglianese, L., et al, "An Avionics Domain-Specific Software Architecture Program," Proceedings of the DARPA Software Technology Conference 1992, Los Angeles, CA, April 1992.
9. Computer Command and Control Company, Final Report for Contract No. N60921-90-C-0298, "Software Intensive Systems Reverse Engineering," April, 1992.
10. Computer Command and Control Company, "Software Specification Assistant" Guides: Status Manager and Step-by-Step Guide, Document Manager Guide, Evaluation Guide and Installation Guide, Contract N00014-91-C-0160, December 1992.
11. Cvetanovic, Z., "The Effects of Problem Partitioning, Allocation and Granularity on the Performance of Multiple-Processor Systems," *IEEE Transactions on Computers*, Vol. C-36, No. 4, April 1987.
12. DeMarco, T. "Structured Analysis and System Specification," Prentice-Hall, Inc., Englewood Cliffs, NJ, 1987.
13. Digital Equipment Corporation, "Guide To DECdesign," Maynard, MA, May 1990.
14. Digital Equipment Corporation, "A Tool Integration Standard," ANSI X3H 4, Information Resource Dictionary System, ATIS, DEC Maynard, MA, February 1990.
15. Digital Equipment Corporation, "DECdesign: User's Guide," AA-PABRB-TE, Maynard, MA, May 1991.
16. Digital Equipment Corporation, "Digital: CDD/Repository: Using CDD/Repository on VMS Systems," DEC, Maynard, MA, Part No. AA-P51KA-TE, Oct. 1991.
17. Digital Equipment Corporation, "Digital: CDD/Repository: Architecture Manual," DEC, Maynard, MA, Part No. AA-PJ1JA-TE, Oct. 1991.
18. Digital Equipment Corporation, "Digital: CDD/Repository: Callable Interface Manual," DEC, Maynard, MA, Part No. AA-PJ1LA-TE, Oct. 1991.
19. DOD, DOD-STD-2167A: "Defense System Software Development," September 1988.
20. DOD, "Military Standard Software Development and Documentation (Draft)," DOD Harmonization Working Group, December 1992.
21. Donnelly, C. and R. Stallman, "BISON, The YACC Compatible Parser Generator" Free Software Foundation, Cambridge, MA 02134, 1990.
22. Evans, A. and Butler, K. J., "Descriptive Intermediate Attributed Notation for Ada Reference Manual," TL-83-4, Tartan Labs, Pittsburgh, PA 1983.
23. Foreman, J. "STARS: State of the Program," STARS '92 Conference, 1992, pp. 20-41.
24. Goldman, S.L. and Roger N. Nagel, "Management, Technology, and Agility: the Emergence of a New Era in Manufacturing," *International Journal Technology Management*, Vol. 8, No. 1/2, pp. 18-38, 1993.
25. Hammer, M. & J. Champy, "The Reengineering Corporation, A Manifesto For Business Revolution," HarperCollins, 1993.
26. Hayes-Roth, F., et. al, "Domain-Specific Software Architectures: Distributed Intelligent Control and Management (DICAM) Applications and Development Support Environment," Proceedings of the DARPA Software Technology Conference 1992, Los Angeles, CA April 1992.
27. Iacocca Institute, "Benchmark Agility," Lehigh University, Bethlehem, PA, 1992.
28. Kimball, J. and A. Srivastava, "AAA 1.0: An Experimental Notation For Engineering Processes," Honeywell Systems and Research Centers, Technical Report CS-R92-013, July 1992.

29. Kimball, J. and K. Thelen "AAA Example: Structured Technical Review," Honeywell Systems and Research Center, June 1993.
30. Lock, E., N. Prywes, and S. Andrews, "Case For Development And Re-engineering Of Real-time Distributed Applications," Fourth International Conference, Software Engineering and Its Applications, Toulouse, France, December 9-13, 1991.
31. Mettala, E. and M. Graham, "The Domain Specific Software Architecture Program," Proceedings of the DARPA Software Technology Conference 1992, Los Angeles, CA, April 1992.
32. Meyers, B., "The State of Art in Visual Programming and Program Visualization," Technical Report CMU-CS-88-114, Carnegie Mellon University, February 1988.
33. Naval Sea Systems Command, PMS 412, User Handbook for CMS-2 Compiler, NAVSEA 0967-LP-598-8020, 30 March 1990.
34. Naval Sea Systems Command, PMS 412, Program Performance Specification for CMS-2 Compiler, NAVSEA 0967-LP-598-9020 30, March 1990.
35. Nielsen, J., "Noncommand User Interfaces," *CACM* (36), No. 4, April 1993, pp. 83-99.
36. Prywes, N., E. Lock, and X. Ge, "Automatic Abstraction of Real-Time Software and Re-Implementation in Ada," Proc of Tri-Ada '91, October 21-25, 1991.
37. Prywes N., I. Lee "Integration of Software Specification, Reuse and Reengineering," Computer Command and Control Company, Philadelphia, PA, 19103, June 1993.
38. PTECH Design and PTECH Code, Release 3.5, Tool User's Guide," Associative Design Technology, Mar, 1992.
39. Selby, R., A.A. Porter, D.C. Schmidt, J. Barney, "Metric Driven Analysis and Feedback Systems for Enabling Empirically Guided Software Development," Proc. 13th International Conference on Software Engineering, 1991.
40. SPC "Domain Engineering Guidebook," Technical Report SPC-92019-CMC, Software Productivity Consortium, December 1992.
41. Srivasta, A. "Strategy For Process Programming Using Agents, Activities And Artifacts (AAA)," Proc. of STARS '92 Conference, December 1992.
42. Stewart, T., "Reengineering: The Hot New Managing Tool," *Fortune*, pp 41-48, August 23, 1993.
43. Sun-Joo Shin, "Valid Reasoning and Visual Representation." PhD thesis, Stanford University, August 1991.
44. Ward, P. and S. Mellor, Structured Design, Yourdon Press, Englewood Cliffs, NJ, 1972, 2nd ed.
45. Wileden, J. A.L. Wolf, W.R. Rosenblatt and P.L. Tarr. "Specification Level Interoperability," *CACM*, 34(5) May 1991.

A HIGH-SPEED LINEAR ALGEBRA LIBRARY WITH AUTOMATIC PARALLELISM

Michael L. Boucher
Dakota Scientific Software, Inc.
501 East Saint Joseph Street
Rapid City, SD 57701-3995

scisoft@well.sf.ca.us
(605) 394-1256 fax
(605) 394-9257 voice

ABSTRACT

Parallel or distributed processing is key to getting highest performance from the latest generation of high-performance workstations. However designing and implementing efficient parallel algorithms is difficult and error-prone. It is even more difficult to write code that is both portable to and efficient on many different computers. Finally, it is harder still to satisfy the above requirements and include the reliability and ease of use required of commercial software intended for use in a production environment. As a result, the application of parallel processing technology to commercial software has been extremely small even though there are numerous computationally demanding programs that would significantly benefit from application of parallel processing. This paper describes DSSLIB, which is a library of subroutines that perform many of the time-consuming computations in engineering and scientific software. DSSLIB combines the high efficiency and speed of parallel computation with a serial programming model that eliminates many undesirable side-effects of typical parallel code. The result is a simple way to incorporate the power of parallel processing into commercial software without compromising maintainability, reliability, portability, or ease of use. This gives significant advantages over less powerful non-parallel entries in the market.

INTRODUCTION

Designing and implementing a parallel algorithm in a way that is both portable and efficient on a wide range of hardware and software configurations is a task that is sufficiently difficult and time-consuming that it is rarely done. Even when developing codes that require only a moderate amount of optimization, it is common to use techniques that are specific to a particular machine and that are not easily portable to other hardware or software architectures. As a program becomes more closely tied to a specific environment it requires more extensive changes in order to adapt to changes in the environment. A predictable undesirable result of writing parallel programs in a way that binds them to a specific hardware or software environment is that the "dusty deck" codes of the future will be even more difficult to use and maintain than the dusty deck codes with which we have to deal today. A predictable result of the pace of technology change is that codes written today will require more frequent updates to adapt, customize, and optimize them for the latest computer system. Therefore, we can expect the maintenance phase to get even more expensive if we continue present approach of requiring that programmers customize programs for a specific environment in order to get acceptable speed. This paper first describes a method of implementing parallelism that avoids many of the problems that inhibit writing portable parallel code, then describes a library of parallel linear algebra subroutines that uses that approach.

Writing portable parallel codes is made difficult by a combination of factors listed below and expanded upon in the paragraphs that follow.

1. It is much more difficult to adequately test parallel code due to nondeterminism in the order in which operations are performed and a lack of good analysis tools.
2. Faults and events, such as division by zero or overflow, tend to be masked, reported incorrectly, or reported inconsistently from one run to the next.
3. Parallelization can slightly change the numerical properties of a given method.

4. Wide variations in the performance characteristics of different parallel and distributed architectures make it difficult to write a single code that is efficient on a range of machine types.
5. Dependency on the run-time environment makes it difficult to write a single code that is efficient on a single machine type under varying system loads.

Parallel algorithms, especially those that use medium- and coarse-grain parallelism, are almost intrinsically nondeterministic in their execution. For example, the order in which synchronizing semaphores are accessed can cause significant changes in the internal behavior of a program. It is possible that some, but not all, access patterns to semaphores or other global data will work properly. It is therefore possible that a bug will go undetected through even extensive structured testing. Few of the standard analysis tools such as static analyzers have adequate support for parallel programs and so those tools are not as helpful with parallel programs as they are with serial programs.

Virtually none of the available parallel systems report faults or events that take place on a parallel CPU. For example, a division by zero or an overflow on a parallel CPU will generally go undetected by the host processor. Virtually all of those systems that do report faults or events back to the host processor do so in a nondeterministic manner. This can significantly complicate the task of debugging.

Parallelism may change some of the numeric properties of a code. There are many well-known examples of this effect; one obvious example is that splitting a summation in different ways can generate different results.

Parallel and distributed architectures are available for all classes of machines ranging from PCs to supercomputers and each of these architectures has widely varying performance characteristics. The parallelism on a given computer system may be fine-, medium, or coarse-grain parallelism, or it may be any combination of those three models. Fine-grain parallelism can take the form of an instruction pipeline or independent functional units in a single CPU. There may also be multiple processor types in a single computer, for example independent CPU and I/O processors or a CPU and an FPU. Medium-grain parallelism is typically loop-level parallelism between several processors with a shared memory. Coarse-grain parallelism can occur between any two processors regardless of whether they share a common memory. All of this variation makes it difficult to design a code that will run well on many or all of the architectures.

Finally, variations in the run-time environment can make it very difficult to write code that is efficient even on a single machine type but under varying work loads. For example, distributing a computation across workstations in a cluster can be done efficiently when the workstations are available and the network is lightly loaded, but inefficient when the workstations are busy or if the network is heavily loaded. Finally, changes in problem size can significantly change the performance characteristics of a particular parallel algorithm.

BACKGROUND / EXISTING APPROACHES

We start by considering the systems for parallel and distributed processing that are widely available today. They appear to fall into one of three categories: remote procedure calls, subroutine libraries that provide parallelism primitives, and pre-parallelized subroutine libraries. The systems that we consider as the representatives of each of these categories are UNIX™ RPCs as implemented by Sun Microsystems [8], Parallel Virtual Machine, and LAPACK [1].

RPCs are a mechanism in which a UNIX programmer can run a procedure on a remote machine using the simple semantics of an ordinary procedure call. When invoked, a synchronous RPC transmits the arguments to a remote machine which then executes the procedure and returns the results. In this way, RPCs provide distributed processing. A layer called XDR tries to hide from the programmer machine-specific details about byte ordering, word length, and so forth by doing some of the data conversion necessary to make the data in the arguments understandable to the remote machine and to make the result from the remote machine understandable to the host. An asynchronous RPC is similar to a synchronous RPC except that the host does not wait for a result from a remote machine after initiating a remote procedure. After initiating a remote procedure on one machine, the host

may make one or more other asynchronous RPCs and in this way the host achieves parallel processing. RPCs are supported by `rpcgen`, which allows a programmer to create RPC templates relatively easily. The strengths of the RPC are its ease of invocation using standard procedure call semantics and its relatively easy portability among UNIX operating environments.

The standard form of RPC/XDR has many drawbacks. First, XDR has a clear bias towards C programs running on 32-bit machines with IEEE floating point arithmetic and it has poor support for data types that are not common on this configuration. For example, FORTRAN's complex data type (a data type not available in C), double precision floating point on a Cray (a 64-bit machine that does not use IEEE arithmetic), and BCD (often supported on IBM mainframes and 80x87 math coprocessors) are poorly handled by the standard XDR. While there is some portability among UNIX operating environments, there is essentially no hope of easily porting an RPC-based program to a non-UNIX environment. Synchronous RPCs do not allow parallel processing and the asynchronous RPCs that do allow parallel processing are almost hopelessly difficult to use. Synchronous RPCs are also less portable than asynchronous RPCs. RPCs do not duplicate the machine state of the host machine on the remote machine so that special processing options selected on the host will not operate correctly on the remote machine. For example, if a program sets the IEEE rounding mode on the host then computations on the host will round correctly and computations on remote machines will round incorrectly. Finally, RPCs have no fault tolerance. If a temporary network glitch occurs or if a remote machine crashes while an RPC-based program is running, then the program will hang or crash if the user is lucky, or the program will return the wrong answer with not even a hint of trouble if the user is unlucky.

Parallel Virtual Machine (PVM) is the representative of the class of parallel and distributed processing tools that are characterized by giving the user direct access to parallel and distributed processing primitives such as send, receive, initiate task, synchronize, and so forth. Other systems that fall into this category are Linda, Express, and the tasking mechanism built into Ada. PVM was developed by Dr. Jack Dongarra and his team at Oak Ridge National Laboratory (ORNL). It is a library of subroutines that gives a programmer close control over the parallelism employed by an application. PVM is more portable than RPC because PVM is not tied to a specific operating system. Dongarra and his team are considerably more scientifically oriented than the designers of RPC and so PVM correctly handles data types from languages besides C and machines with configurations besides 32-bit CPUs using IEEE arithmetic. PVM is designed to allow parallel processing in addition to simply the distributed processing capability of synchronous RPCs. Parallel processing with PVM is much easier than with asynchronous RPCs.

PVM is generally superior to RPC, but it has some drawbacks. From the perspective of a computer scientist, the power of PVM comes largely from the degree of control that the programmer can exercise over the process of parallelization. From the perspective of an atmospheric scientist, the problem with using PVM is the degree of control that the programmer must exercise over the process of parallelization. Many of the messy details of interprocessor communication that were concealed with RPCs are now the programmers problem. Another drawback to using PVM is that it requires that PVM-based programs be parallel or distributed programs. PVM-based programs that are developed on a multiprocessor SPARCstation 10™ will run beautifully, in large part due to the extremely fast interprocessor communication that comes with shared memory. PVM-based programs that are run on a network of SPARCstation IPXs will run poorly, in large part due to the extremely slow interprocessor communication that comes with the Ethernet connection. Regardless of the extreme variations in efficiency between these two operating environments, PVM forces the program to behave in exactly the same way in both environments. Finally, PVM is slightly better than RPC at fault tolerance, but not much. If a fault occurs in a network or on a remote machine while a parallel computation is in progress, the application probably will fail.

LAPACK represents the approach of using parallel subroutine libraries. In contrast to PVM, whose subroutines allow the user to define the operations involved in building a parallel application, LAPACK is a library of subroutines that may be supplied to a user after being optimized and parallelized. LAPACK includes subroutines to perform many of the common operations in computational linear algebra including solving systems of linear equations, matrix factorizations, eigensystem solvers, SVD, and similar operations.

Much of LAPACK is built on block operations, meaning that it divides a data set into subblocks that can be processed independently. It then does operations on those blocks. These blocks are then mapped for processing to the resources of a given machine. If there are multiple processors present then the blocks may be mapped to processors. The blocks may also be selected to correspond to the size of a cache for efficient memory access. The standard version of LAPACK as it comes from Oak Ridge National Laboratory is not optimized or parallelized, but the block structure does make it simpler to parallelize than other subroutine libraries that perform similar functions. LAPACK uses a subroutine called ILAENV to help it determine how each subroutine call should be blocked and so it is possible for ILAENV to react to changes in the environment and adapt its parallelism strategy accordingly. A major drawback to LAPACK with respect to its utility as a parallel programming environment is the same as its major strength, which is that the programmer has no concern with or control over the parallel processing. As a result, the programmer has no way to extend the parallel processing to get some capability that is not built into LAPACK.

A PROPOSED SOLUTION: DSSLIB

We have developed a parallelization system named DSSLIB that will avoid many, though not all, of the pitfalls of the available parallel programming systems. In particular, because it is a library, DSSLIB has the drawback present in LAPACK that a user cannot extend it to perform computations that are not built in. DSSLIB is based on a combination of software programs transferred from a variety of US. Government agencies and projects. Some of the software has been in wide use since 1979 while others have been introduced as recently as 1991. Specifically, DSSLIB includes version 1.1 of LAPACK and the latest versions of LINPACK [3] and levels 1, 2, and 3 of the Basic Linear Algebra Subprograms (BLAS) [6, 5, 4]. We intend this system for use in production codes, including commercial software, users who are not sophisticated programmers of parallel or distributed processing machines, and for any user regardless of sophistication who needs a significant speedup in an application but does not have the resources to dedicate to a parallelization effort.

The choice of target users implies that the software must have at least the following characteristics:

1. most or all of the parallelization must be automatic
2. runs correctly and reasonably efficiently in a variety of hardware configurations
3. complete fault tolerance.
4. does not interfere with other software that may be in use, possibly including other parallelization systems
5. compatible with all of the standard tools such as debuggers, profilers, etc.
6. requires no changes to move among many different hardware and software configurations; retains all of the characteristics listed above even as it is being used in a variety of configurations

DSSLIB satisfies the criteria above by presenting to an application a serial programming model even when it is running in parallel. A serial programming model means that DSSLIB appears to an application to be a standard library running on a single CPU. Some of the implications of choosing a serial programming model are:

1. for a given set of data, results will always be exactly the same regardless of how a particular computation was parallelized on a specific run
1. standard tools such as debuggers and profilers continue to work in the same way that they always have
2. IEEE conditions are presented to an application in exactly the same way regardless of whether a computation is performed serially or in parallel
3. DSSLIB always presents signals to an application in the same way every time
4. parallel machines or processors use exactly the same environment as the host machine

Given our choice of target users, one of the most important requirements is that all of the parallelization be automatic. When an application calls one of the parallel subroutines then DSSLIB determines how many processors to use, how to partition the data, and divides the work among the available processors. The number of processors to use is computed based on the size of the computation, expected performance from the network, expected performance from the other processors, and other factors. For a given computation, the number of

processors assigned may vary as an application runs due to changes in the factors that influence the number of processors assigned. However, DSSLIB partitions all computations in a way that guarantees that for a given computation the answer returned from DSSLIB will always be bitwise identical, regardless of the number of processors.

As part of doing the automatic parallelism, DSSLIB records certain performance information about each computation and continuously tunes itself as a program runs. In addition to making the automatic parallelism more efficient, this has the interesting side-effect that large applications will tend to get faster as they run because DSSLIB will have time to learn and adapt to the environment. For example, an application may learn that the network is more lightly loaded than expected and that the cost of communicating with remote processors is less than anticipated. As a result, it may choose to use more processors for future computations than it would have chosen by default. To illustrate this property, we modified the LINPACK 1000x1000 benchmark to solve six linear systems and record six times instead of just one. In that test, DSSLIB solved the sixth linear system 18% faster than it solved the first system because it was able to apply to the sixth computation things that it had learned about the environment while doing the earlier computations.

The fact that the parallelism is automatic allows DSSLIB to be incorporated into production code, even commercial software. While it is possible for a researcher to specify in advance a detailed and possibly very narrow description of the types of problems to be solved, a commercial software package will be presented with a variety of problems or varying sizes and shapes. The researcher writing a specialized code to solve a narrow set of problems may know in advance a close estimate to the optimal number of processors, but a commercial package cannot have such assumptions built in. The researcher may have available the luxury of being able to schedule a block of time on an empty or nearly empty system. A production code may be run in such an environment, but it also needs to work well in a shared environment. The automatic and adaptive parallelism in DSSLIB allows the flexibility required of commercial or production software.

Of course, one of the characteristics of a serial programming model is that a program generally will not be adversely affected by problems in the network or on other computers. To support this aspect of a serial programming model, DSSLIB has been built completely fault tolerant. If one or more parallel machines crash due to problems in hardware, software, or network then DSSLIB will detect the problem and automatically restructure the computation so that it will complete correctly. Further, it will restructure the computation in such a way as to guarantee that the answer from the restructured computation will be bitwise identical to the answer that would have been computed under normal conditions. Of course, failures on the host machine can hurt the application, but this also is consistent with the serial programming model. As with the automatic parallelization above, the compensation for faults or errors is automatic and does not require any special code on the part of the user.

Detected errors or faults, such as IEEE conditions, are presented to an application in exactly the same way every time regardless of whether a computation is performed serially or in parallel. For example, if an application attempts to solve a singular linear system then the subroutine DGESE will divide by zero. (This is standard documented LINPACK behavior, not a DSSLIB problem.) Most other parallel systems will not return a divide by zero indication to a user's application. DSSLIB will return divide by zero or any other condition to a user's application exactly as if it had been run on a single CPU. Also, DSSLIB always presents multiple signals to an application in the same order every time. Consider a computation that will be performed in parallel in which one parallel machine will divide by zero and another will get an overflow. DSSLIB guarantees that those signals will always be presented to the user's application in the same order every time, just as they would be if the computation were performed on a single CPU. DSSLIB has no race conditions common in other parallel systems.

Parallel machines or processors use exactly the same environment as the host machine. For example, if an application changes the IEEE rounding mode to round towards zero instead of round to nearest then all parallel machines or processors will round towards zero. Other parallel processing systems do not ensure that parallel computations are performed in the environment requested by an application.

RESULTS

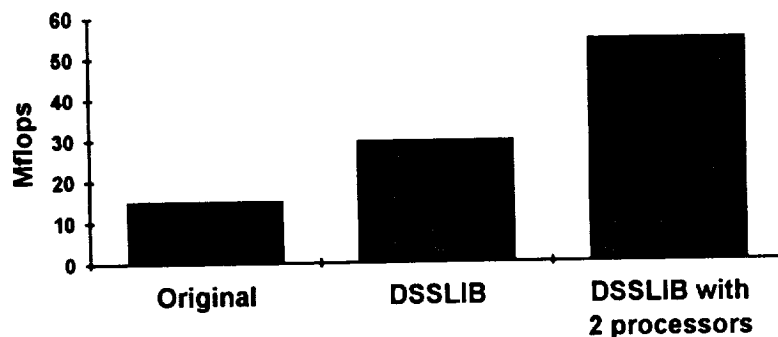
Of course, the acid test of any parallel or distributed system is speed. If it is not fast then none of its other characteristics are particularly interesting. It is even more helpful if the system is fast on real applications, rather than running well only on a selected set of benchmarks. DSSLIB is fast. Measurements on applications with run times ranging from 90 seconds to 12 hours shows that it delivers a pleasing level of performance for a reasonable variety of applications. Further, DSSLIB satisfies the requirement that it run well in a variety of hardware and software environments with no changes required of the user.

The hardware configurations in which DSSLIB was tested for this paper include both parallel and distributed processing where processors are defined to be parallel if they share a common memory. The shared memory machine used for this paper was a dual processor SPARCstation 10 with a 40 MHz SuperSPARC processor. Processors are defined to be distributed if they do not share a common memory but are linked via some network such as Ethernet or FDDI. Distributed processing machines used for this paper were single processor SPARCstation IPXs linked by a network. Networks used were the standard Ethernet and the SBUS FDDI product from Network Peripherals.

Both fine- and coarse-grain parallelism was tested. Fine-grain parallelism was done by making best use of the parallelism between the floating point and integer CPUs, and also between the independent add and multiply units in the floating point unit. On the SS10, additional fine-grain parallelism was measured by taking advantage of the multiple instruction per cycle capability, though this was limited by the fact that floating point instructions launch one at a time.

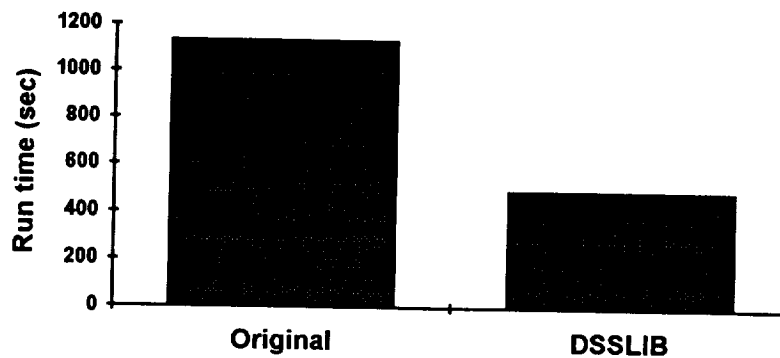
The software on which DSSLIB was tested included a matrix multiplication benchmark, a small image processing program written in IDL¹ that has a run time of 90 seconds, an artificial neural network written in FORTRAN 77 with run-times of 19 and 58 minutes for two data sets, and a discrete ordinate radiative transfer program [7] written in FORTRAN 77 with a run-time of 12 hours.

We ran the matrix multiplication benchmark on a single-CPU SPARCstation 10 model 40, then again on a dual processor machine. This benchmark simply generates 400x400 matrices and computes $\alpha AB + \beta C \rightarrow C$. The comparison below shows the speed of the standard form of DGEMM from netlib and the speed of the same subroutine from DSSLIB. The DSSLIB subroutine is timed on one and two processors where the single processor run takes advantage of only fine-grain parallelism and the two processor run takes advantage of all levels of parallelism.

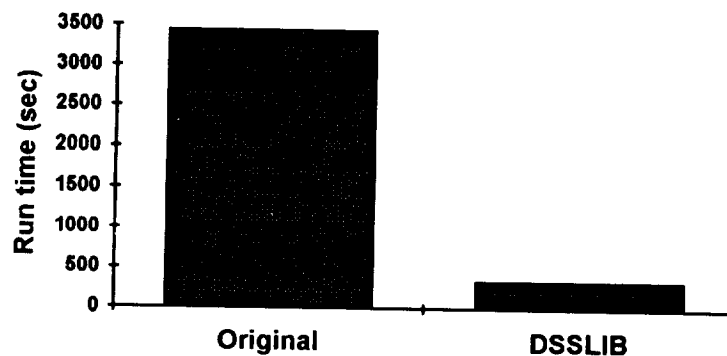


¹IDL is an interpreted data modeling language from Research Systems, Inc. It appears to a user to be nearly identical in most ways to PV-WAVE from Visual Numerics, and the results reported for IDL are virtually identical to the results of similar experiments with PV-WAVE.

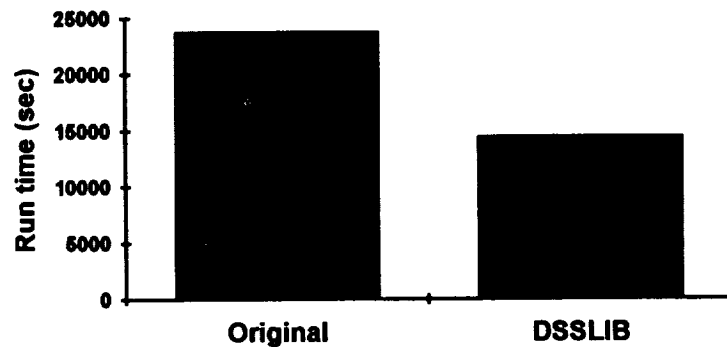
The neural network tested was part of a NASA project to classify cloud formations extracted from satellite data. The large Landsat data set was first reduced to a set of feature vectors extracted during a preprocessing phase and these feature vectors were used as input to the neural network. The neural network then classified the cloud formations in the image according to the information found in the feature vectors. The graph below shows the results obtained with a small data set. This data set is sufficiently small that it runs on a single CPU and so uses only fine-grain parallelism. The graph shows wall clock run time, so smaller values indicating shorter run times are better.



As one can clearly see from the graph above, the introduction of DSSLIB had a significant positive effect on the performance of the program, even for modest-sized data sets. Based on these results, Logar and Corwin decided to eliminate the data reduction in the preprocessing phase and send the raw satellite data directly to the neural network. In its present form, the neural network is limited by the size of the main memory on the Sun workstation rather than by the speed of the Sun CPU. The results below are for two Sun IPX workstations linked by an SBUS FDDI card from Network Peripherals.



DISORT is an application available used by NASA Goddard Space Flight Center to compute the thermal budget of a two-dimensional multi-layer region of the Earth's atmosphere. Each individual horizontal layer is required to be homogeneous, but different layers may have different characteristics. This program is dominated by eigenvalue computations done with modified subroutines from EISPACK. It also uses a significant amount of CPU time on LINPACK and various matrix algebra computations from the BLAS libraries. Because the EISPACK subroutines had been modified, there is no safe way to replace them with LAPACK subroutines, which is what one would usually do. However it is possible to insert a few BLAS calls into the eigenvalue subroutines and other places in the program. We did this in a way that provided some speed improvement, but did not compromise the accuracy or portability of the program. The results are shown in the following graph.



As one would expect from a program dominated by eigenvalue computation, parallelism provides significant speedup, but the improvement is not a factor of two for two processors. Nevertheless, DSSLIB cut three hours from a seven hour run using two processors. The customer is now able to make runs that would have been prohibitively expensive before these changes.

SUMMARY

DSSLIB is a library built on a parallel processing system that presents to an application a serial programming model. This serial programming model simplifies the development of a parallel application because it hides from the programmer the difficult details of parallelization. It also allows DSSLIB to be incorporated into production or commercial software because it is able to adapt to the variety of configurations and environments in which such software is used.

BIBLIOGRAPHY

- Anderson, E., Z. Bai, C. Bischof, J.W. Demmel, J.J. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. LAPACK User's Guide, Society of Industrial and Applied Mathematics, Philadelphia, Pa., 1992.
- Bernard G. et. al. Primitives for Distributed Computing in a Heterogeneous Local Area Network Environment. *IEEE Transactions on Software Eng.*, 15 (12), December 1989, pp 1567-1578.
- Dongarra, J.J., C.B. Moler, J.R. Bunch, G.W. Stewart. LINPACK User's Guide, Society of Industrial and Applied Mathematics, Philadelphia, Pa., 1979.
- Dongarra, J.J., J. DuCroz, I. Duff, and S. Hammarling. A Set of Level 3 Basic Linear Algebra Subprograms, *ACM Transactions on Mathematical Software*, 16 (1990), pp 1-17.
- Dongarra, J.J., J. DuCroz, S. Hammarling, and R.J. Hanson. An Extended Set of FORTRAN Basic Linear Algebra Subprograms, *ACM Transactions on Mathematical Software*, 14 (1988), pp 1-17.
- Lawson, C.L., Hanson, R.J., Kincaid, D., and Krogh, F.T. Basic Linear Algebra Subprograms for FORTRAN Usage, *ACM Transactions on Mathematical Software*, 5 (1979), pp 308-323.
- Stamnes, K., Tsay, S., Wiscombe, W., Jayaweera, K. Numerically Stable Algorithm for Discrete-Ordinate-Method Radiative Transfer in Multiple Scattering and Emitting Layered Media. *Applied Optics*, 27 (1988), pp 2502-2509.
- Sun Microsystems, Inc. Network Programming Guide. Revision A, March 1990, pp 33-167.

omit

INFORMATION MANAGEMENT

HIGH-SPEED DATA SEARCH

James Driscoll
Department of Computer Science
University of Central Florida
Orlando, Florida 32816, USA

ABSTRACT

The high-speed data search system developed for KSC incorporates existing and emerging information retrieval technology to help a user intelligently and rapidly locate information found in large textual databases. This technology includes: natural language input; statistical ranking of retrieved information; an artificial intelligence concept called semantics, where "surface level" knowledge found in text is used to improve the ranking of retrieved information; and relevance feedback, where user judgments about viewed information are used to automatically modify the search for further information. Semantics and relevance feedback are features of the system which are not available commercially. The system further demonstrates a focus on paragraphs of information to decide relevance; and it can be used (without modification) to intelligently search all kinds of document collections, such as collections of legal documents, medical documents, news stories, patents, and so forth. The purpose of this paper is to demonstrate the usefulness of statistical ranking, our semantic improvement, and relevance feedback.

INTRODUCTION

Locating information using large amounts of natural language documents (text) is an important problem. Examples at KSC are searching press releases and numerous other documents to quickly answer media questions, accessing bulky manuals and schematics compactly stored on a CD via a laptop computer, and retrieving digital images by means of their catalog descriptions.

The primary intent of our work has been to provide convenient access to information contained in the numerous and large public information documents maintained by Public Affairs at NASA Kennedy Space Center (KSC). The documents maintained by Public Affairs at NASA KSC consist of press releases, and other printed information created at KSC, and other NASA offices using various wordprocessors. There are also documents from outside contractors, such as Rockwell, which produces the "NASA National Space Transportation System Reference" more often called the "shuttle manual." During a launch at KSC, about a dozen NASA employees access these printed documents to answer media questions. The planned document storage for NASA KSC Public Affairs is around 300,000 pages (approximately 900 megabytes of disk storage).

Current commercial text retrieval systems focus on the use of keywords to search for information. These systems typically use a Boolean combination of keywords supplied by the user to retrieve documents. In general, the retrieved documents are not ranked in any order of importance, so every retrieved document must be examined by the user. This is a serious shortcoming when large collections of documents are searched.

The QA system is a high-speed data search system developed jointly by NASA KSC, the University of Central Florida, and Florida High Technology and Industry Council. It is a statistically based text retrieval system which ranks retrieved documents according to their statistical similarity to a user's request. Statistically based systems provide many advantages over traditional Boolean retrieval methods, especially for users of such systems, mainly because they allow natural language input. These systems have been a research success for over twenty years [9]. However, the transfer of this retrieval technique into large operational systems has been very slow because, until recently, there was no evidence that statistical ranking could be done in real-time on large document collections [4]. There are only three commercial systems in the United States which allow natural language input and perform statistical ranking of retrieved information [2].

The QA System incorporates two other features which are not available in any commercial text retrieval system, but have been shown to dramatically improve the statistical ranking of retrieved information. The first is an artificial intelligence concept called semantics, where "surface level" knowledge found in text is used to improve the ranking of retrieved information. The second is relevance feedback, where user judgments concerning viewed information are used to automatically modify the search for more information.

The QA System is very close to being a commercial product. It has been used to participate in a (first) Text Retrieval Conference (TREC-1) managed by the National Institute of Standards and Technology (NIST). Our participation in TREC-1 was funded by the Defense Advanced Research Projects Agency (DARPA). Participation in TREC-1 has enabled the QA System to be tested in an environment other than answering questions, and applied to databases other than aerospace text collections [3].

Conventional information retrieval using statistical ranking is demonstrated first in this paper. Demonstrations of improved statistical ranking due to the use of semantics within the QA System are then presented for comparison. This is followed by a demonstration of relevancy feedback within the QA System. In all demonstrations, the focus on paragraphs of information for retrieval will be evident. Finally, the issues of platforms and high-speed for the QA System are discussed in the Conclusion.

CONVENTIONAL INFORMATION RETRIEVAL

Finding relevant text and ranking the retrieved documents is not new and there are commercial systems which already perform this activity; we mention here an example of ranked, relevant text retrieval. For a demonstration to NASA KSC, the 1000 page shuttle manual was used by considering each paragraph of the manual as a document. This resulted in a collection of 5143 documents. A commercial hypertext IR system called SPIRIT [11] was used to automatically index the collection and provide natural language access. SPIRIT is a mainframe system. Running on an IBM 4381, SPIRIT required three and one-half hours of clock time to index the collection of 5143 documents.

Figure 1 is a screen generated by SPIRIT for asking the natural language query

What are the dimensions of the cargo area in the shuttle?

Figure 2 is a screen generated by SPIRIT revealing a ranked list of 245 relevant documents with CLASS 1 being the most relevant. Figure 3 is a screen generated by SPIRIT revealing the first document in CLASS 6, which contains the answer to the query. This paragraph was found by reading the single paragraph in CLASS 1 first, then the single paragraph in CLASS 2, and so on until the answer was read in the tenth paragraph.

NATURAL LANGUAGE QUERY ON THE SHUTTLE BASE

<1>: What are the dimensions of the cargo area in the shuttle?

EMPTY WORDS: What, are, the, of, the, in, the.

KEYWORDS: dimensions, cargo, area, shuttle.

Figure 1. Natural Language Query to the SPIRIT System.

CLASSES	NB DOCS	KEYWORDS
1	1	dimensions, cargo, shuttle.
2	1	cargo, area, shuttle.
3	1	dimensions, area.
4	2	dimensions, shuttle.
5	4	cargo, area.
6	30	cargo, shuttle.
7	12	area, shuttle.
8	7	dimensions.
9	40	cargo.
10	147	area.
BOTTOM OF LIST		

Figure 2. Document Classes Generated by the SPIRIT System.


```

DOC 0005 BASE : doc 0005NCP:0/CPI:1/NBI:1+18 1K/1K
IDENTIFIER. : doc 0005
TEXT..... :

The shuttle will transport cargo into near Earth orbit 100 to 217 nautical miles (115 - 250
statute miles) above the Earth. This cargo (called payload) is carried in a bay 15 feet in
diameter and 60 feet long.
BOTTOM OF DOCUMENT

INFORMATIONAL PAGE 1/1
WHAT DO YOU WANT TO DISPLAY?
> OR RETURN,<,>,<<,DOC,END,DDQ,(?):

```

Figure 3. Document Display by the SPIRIT System.

Note that performance in this Question/Answer environment is measured by counting how many documents were examined to find the document containing the answer. This is not the usual way of measuring the performance of IR systems, but it is very appropriate for a Question/Answer environment.

The underlying principles and algorithms of automated IR systems like SPIRIT are well-known. Terms used as document identifiers are keywords modified by various techniques such as stop lists (removal of useless or empty words), stemming, synonyms, and query reformulation. Here, we present basic concepts associated with the calculation of weighting factors.

The calculation of the weighting factor (w) for a term in a document is a combination of term frequency (tf), document frequency (df), and inverse document frequency (idf). The basic term definitions are as follows:

$$\begin{aligned}
 tf_{ij} &= \text{number of occurrences of term } T_j \text{ in document } D_i \\
 df_j &= \text{number of documents in a collection which contain } T_j \\
 idf_j &= \log\left(\frac{N}{df_j}\right), \text{ where } N = \text{total number of documents} \\
 w_{ij} &= tf_{ij} \cdot idf_j.
 \end{aligned}$$

When an IR system is used to query a collection of documents with t terms, the system computes a vector Q equal to $(w_{q1}, w_{q2}, \dots, w_{qt})$ as the weights for each term in the query. The retrieval of a document with vector D_i equal to $(d_{i1}, d_{i2}, \dots, d_{it})$ representing the weights of each term in the document is based on the value of a similarity measure between the query vector and the document vector. A common similarity function which normalizes the the similarity coefficient in case of different document sizes is the following:

$$\text{sim}(Q, D_i) = \frac{\sum_{j=1}^t w_{qj} \cdot d_{ij}}{\sqrt{\sum_{j=1}^t d_{ij}^2}} \quad (1)$$

It is important to note that the calculation of a similarity coefficient for each document and the ranking of the documents relevant to a query is rather time consuming. This is due to the summations that occur in the above formula and the fact that every document that has a term in common with a given query must be considered. The main problem with text retrieval using statistical ranking has been the time required to produce the document ranking given a query. Consequently, query response time has been typically slow.

SEMANTIC APPROACH

Although the basic statistical ranking approach (as demonstrated by SPIRIT) has shown some success in regard to natural language queries, it ignores some valuable information. We now know that these systems can be further improved by imposing a semantic data model upon the "surface level" knowledge found in text.

Semantic Modeling

Semantic modeling was an object of considerable database research in the late 1970's and early 1980's [1]. Essentially, the semantic modeling approach identified concepts useful in talking informally about the real world. These concepts included the two notions of entities (objects in the real world) and relationships among entities (actions in the real world). Both entities and relationships have properties.

The properties of entities are often called attributes. There are basic or surface level attributes for entities in the real world. Examples of surface level entity attributes are Size, Color, and Position. These properties are prevalent in natural language. For example, consider the phrase "large, black book on the table," which indicates the Size, Color, and Position of a book.

In linguistic research, the basic properties of relationships are discussed and called thematic roles. Thematic roles are also referred to in the literature as participant roles, semantic roles, and case roles. Examples of thematic roles are Beneficiary and Time. Thematic roles are prevalent in natural language, they reveal how sentence phrases and clauses are semantically related to the verbs in a sentence. For example, consider the phrase "purchased for Mary on Wednesday" which indicates who benefited from a purchase (Beneficiary) and when a purchase occurred (Time).

Consider the following query:

How long does the payload crew go through training before a launch?

The basic statistical approach dismisses the following words in the query as empty: "how", "does", "the", "through", "before", and "a". Some of these words contain valuable semantic information. The following list indicates some of the thematic roles triggered by a few of the words in the above query:

long ⇒ Duration, Time
 through ⇒ Location/Space, Motion With Reference To Direction, Time
 before ⇒ Location/Space, Time

As another example, consider the query in Figure 1:

What are the dimensions of the cargo area in the shuttle?

The keyword "dimensions" indicates the attribute General Dimensions and the keyword "area" indicates both the thematic role Location/Space and the attribute General Dimensions. It would be reasonable to expect that the document that answers this query would have words in it that fall in the category of General Dimensions.

The primary goal of the QA System has been to detect thematic and attribute information contained in natural language queries and documents. When the information is present, the system uses it to help find the most relevant paragraph to a query. In order to use this additional information, the basic underlying concept of text relevance was modified. The major modifications include the addition of a lexicon with thematic and attribute information, and a modified computation of the similarity measure given in (1).

The Semantic Lexicon

The QA System uses a thesaurus as a source of semantic categories (thematic and attribute information). For example, Roget's Thesaurus contains a hierarchy of word classes to relate word senses [5]. For our research, we have selected several classes from this hierarchy to be used for semantic categories. We have defined thirty-six semantic categories as shown in Figure 4.

In order to explain the assignment of semantic categories to a given term using Roget's Thesaurus, consider the brief index quotation for the term "vapor":

vapor		
n.	fog	404.2
	fume	401
	illusion	519.1
	spirit	4.3
	steam	328.10
	thing imagined	535.3
v.	be bombastic	601.6
	bluster	911.3
	boast	910.6
	exhale	310.23
	talk nonsense	547.5

Thematic Role Categories	Attribute Categories
Accompaniment	Color
Amount	External and Internal Dimensions
Beneficiary	Form
Cause	Gender
Condition	General Dimensions
Comparison	Linear Dimensions
Conveyance	Motion Conjoined with Force
Degree	Motion in General
Destination	Motion with Reference to Direction
Duration	Order
Goal	Physical Properties
Instrument	Position
Location/Space	State
Manner	Temperature
Means	Use
Purpose	Variation
Range	
Result	
Source	
Time	

Figure 4. Thirty-Six Semantic Categories.

The eleven different meanings of the term "vapor" are given in terms of a numerical category. We have developed a mapping of the numerical categories in Roget's Thesaurus to the thematic role and attribute categories given in Figure 4. In this example, "fog" and "fume" correspond to the attribute State; "steam" maps to the attribute Temperature; and "exhale" is a trigger for the attribute Motion with Reference to Direction. The remaining seven meanings associated with "vapor" do not trigger any thematic roles or attributes. Since there are eleven meanings associated with "vapor," we indicate in the lexicon a probability of 1/11 each time a category is triggered. Hence, a probability of 2/11 is assigned to State, 1/11 to Temperature, and 1/11 to Motion with Reference to Direction. This technique of calculating probabilities is being used as a simple alternative to a corpus analysis. It should be pointed out that we are still experimenting with other ways of calculating probabilities.

Extended Computation of the Similarity Measure

The probabilistic details of a semantic lexicon and the computation of semantic weights can be found in [13]. A detailed explanation of the manner in which the QA System combines semantic weights and keyword weights can be found in [12].

Essentially we treat semantic categories like indexing terms, and the probabilities introduced by a semantic lexicon mean that the frequency of a category in a document becomes an expected frequency and the presence of a category in a document becomes a probability for the category being present. This means that the document frequency for a category becomes an expected document frequency, and this enables an inverse document frequency to be calculated for a category.

So the computation of a similarity coefficient as shown in (1) can be used, but now the summations in the formulas include semantic categories in the documents as well as terms in the documents. In other words,

$$sim(Q, D_i) = \frac{\sum_{j=1}^s w_{qj} \cdot d_{ij} + T \sum_{j=s+1}^{s+s} w_{qj} \cdot d_{ij}}{\sqrt{\sum_{j=1}^s d_{ij}^2 + B \sum_{j=s+1}^{s+s} d_{ij}^2}} \quad (2)$$

where $s = 36$ is the number of semantic categories, and T and B are scaling factors for adjusting the blend.

SEMANTIC IMPROVEMENT

The QA System has demonstrated a noticeable semantic improvement using the similarity function in (2). Consider the same document collection and natural language query shown in the commercial system example of Figures 1, 2, and 3. Using the commercial system SPIRIT, ten paragraphs were read in order to find the answer to the following query:

What are the dimensions of the cargo area in the shuttle?

Considering the QA System, Figure 5 is a screen generated for asking this same natural language query. Figure 6 is a screen generated by the QA System graphically showing to the user the importance of the keywords found in the query. Figure 7 is a screen generated by the QA System graphically showing to the user the importance of semantic information found in the query. Notice the "importance" of the semantic category General Dimensions in the screen shown in Figure 7. This long bar means that the semantic category General Dimensions is present in the query and there are very few documents retrieved (using keywords) having this type of semantic content. Hence, the importance of the category.

Finally, Figure 8 is a screen generated by the QA System revealing the second paragraph found by proceeding through the ranked list of documents retrieved by the QA System for this query. The semantic information found in the query and displayed in Figure 7 is the reason the QA System ranked the answering paragraph second instead of tenth as did the SPIRIT system. Notice that the answering document in Figure 8 has several words in it which trigger the semantic category General Dimensions. We have lots of data like this and several technical papers which reveal a significant performance improvement due to semantic modeling in the NASA KSC Question/Answer environment.

For another example of semantic improvement, consider the shuttle manual and the query:

How fast does the orbiter travel on orbit?

This query is interesting for two reasons. One is that the words "orbiter" and "orbit" are rather frequent words in the shuttle manual so lots of paragraphs are retrieved. The other reason is that the word "fast" is used for reference to velocity or speed.

Figure 9 shows the number of paragraphs one must read to find a particular answering paragraph to this query for both a small and large collection of documents. In the small collection, the word "fast" does not occur at all and for the large collection, the word "fast" never occurs in an answering paragraph. Consequently, keyword only statistical ranking is never very good. But by using semantics, the word fast causes a similarity to paragraphs using the words velocity or speed. Consequently, semantics improves the statistical ranking of an answering paragraph. Different blends of keywords and semantics are shown using the similarity function in (2).

RELEVANCE FEEDBACK

It has been pointed out that conventional IR systems have a limited recall [6]; only a few relevant documents are retrieved in response to user queries if the search process is based solely on the initial query. This indicates a need to modify (or reformulate) the initial query in order to improve performance. It is customary to search the relevant documents iteratively as a sequence of partial search operations. The results of earlier searches can be used as feedback information to improve the results of later searches. One possible way to do this is to ask the user to make a relevance decision on a certain number of retrieved documents. Then this relevance information can be used to construct an improved query formulation and recalculate the similarities between documents and query in order to re-rank them. This process is known as relevance feedback [7,8,9,10] and it has been shown experimentally to improve the performance of the retrieval system.

The basic assumption behind relevance feedback is that, for a given query, documents relevant to it should resemble each other in a sense that they have reasonably similar keyword vectors. This implies that if a retrieved document is identified as relevant, then the initial query can be modified to increase its similarity to such a relevant document. As a result of this reformulation, it is expected that more of the relevant documents and fewer of the nonrelevant documents will be extracted.

The automatic construction of an improved query is actually straightforward, but it does increase the complexity of the user interface and the use of the retrieval system, and it can slow down query response time. Essentially, the terms and semantic categories for documents viewed as relevant to a query can be used to modify the weights of terms and semantic categories in the original query. A modification can also be made using documents viewed as not relevant to a query. Experimental results show a very promising improvement for relevance feedback within the QA System.

	Keywords Only	$T - B = 1.10206$ Blend of Keywords and Semantics	$T - B = 8.0$ Blend of Keywords and Semantics
First 26 pages of the shuttle manual (160 documents)	19	4	2
The entire shuttle manual (5143 documents)	145	126	14

Figure 9. Number of paragraphs read to find a particular answering paragraph for:
How fast does the orbiter travel on orbit?

Figure 10 provides an example using the first 26 pages of the shuttle manual and the query:

How fast does the orbiter travel on orbit?

Recall from Figure 9 that 19 paragraphs were read to find an answering paragraph. The document identifiers for these 19 paragraphs are shown in the left column of Figure 11 along with the notes that Document #13 and Document #16 were considered relevant to the original query, and Document #14 answered the query. All the other viewed documents were not relevant to the query.

If relevance feedback is selected within the QA System and the system is told to display two documents and then reformulate the query, then the documents shown in the right column are viewed. Each document viewed must be tagged as relevant or not-relevant. Document #14 shows up earlier in the statistical ranking primarily because Document #13 was tagged as relevant to the original query.

It is interesting to note that if one tags Document #14 (which answers the query) as relevant, then Document #87 is retrieved and it almost exactly answers the query. Document #87 would never be retrieved using just keywords without feedback because it has no keywords in common with the original query. Documents 13, 14, 16, 69 and 87 are shown in Figure 11. The keywords that these documents have in common with the original query are underlined. Clearly, Document 69 is not relevant to the original query.

CONCLUSION: PLATFORMS AND THE ISSUE OF HIGH SPEED

Originally, the QA System was restricted to an IBM compatible PC platform running under the DOS operating system and without the use of any other licensed commercial software such as a DOS extender. The QA System is implemented in Borland C and one version uses B+ tree structures for the inverted files. We felt the speed of the system and its storage overhead was not efficient so a hashing scheme was added to eliminate the use of B+ trees and provide codes for keywords. We expected this second version to have improved indexing time, storage, and retrieval speed.

Experiments revealed that indexing time of the QA System did not improve much. We were not surprised because the QA System is restricted under the PC DOS platform. This platform has a serious memory addressing restriction which results in memory page swapping and this seriously affects the speed of processing, especially during creation of the hashing table and index structures. The improvement in storage, however, was very impressive. It is very much matched to our objective which is to make our storage ratio of indexes to text, around 0.5. This is comparable to the ratio of very efficient, retrieval systems using statistical ranking.

Addressing the high speed issue, we now have the Borland C compiler for OS/2 so we expect to have a very high speed QA System running under OS/2 very soon. We are also in the process of converting the QA System to run in the UNIX environment. Figure 12 reveals achieved and projected run-time performances of the QA System on different operating system platforms. The DOS, B+ tree version of the system is shown in the upper left corner. Below (diagonally) are shown the OS/2, UNIX B+ tree and hashing versions of the QA System for different amounts of RAM. Indexing and typical query response times are shown for both a small (2.4 megabyte) and a large (1.2 gigabyte) document collection. Data for this chart was obtained in part from experiments performed for TREC-1 [3].

160 Documents			
Answer can be found in Document 14, 87			
Keywording		Relevance Feedback (view 2)	
1	69	- 1	69
2	13	- 2	13
3	82	- 3	82
4	15	- 4	107
5	123	- 5	85
6	106	- 6	124
7	85	- 7	16
8	124	- 8	14
9	21	- 9	87
10	23		
11	24		
12	83		
13	31		
14	26		
15	16		
16	84		
17	11		
18	12		
19	14		
:			
:			
never get 87 (no query words in 87)			

Figure 10. Relevance Feedback Improvement for the Query:
How fast does the orbiter travel on orbit?

Document 13
The two <u>orbital</u> maneuvering system engines are used to place the <u>orbiter</u> on <u>orbit</u> , for major velocity maneuvers on <u>orbit</u> and to slow the <u>orbiter</u> for re-entry, called the deorbit maneuver. Normally, two <u>orbital</u> maneuvering system engine thrusting sequences are used to place the <u>orbiter</u> on <u>orbit</u> , and only one thrusting sequence is used for deorbit.
Document 14
The <u>orbiter's</u> velocity on <u>orbit</u> is approximately 25,405 feet per second. The deorbit maneuver decreases this velocity approximately 300 feet per second for re-entry.
Document 16
For deorbit, the <u>orbiter</u> is rotated tailfirst in the direction of the velocity by the primary reaction control system engines. Then the <u>orbital</u> maneuvering system engines are used to decrease the <u>orbiter's</u> velocity.
Document 69
- Atlantis (OV-104), after a two-masted ketch operated for the Woods Hole Oceanographic Institute from 1930-1966, which <u>traveled</u> more than half a million miles in ocean research.
Document 87
Entry interface is considered to occur at 400,000 feet altitude approximately 4,400 nautical miles (5,063 statute miles) from the landing site and at approximately 25,000 feet per second velocity.

Figure 11. Documents 13, 14, 16, 69, and 87. Keywords in common with the original query are underlined.

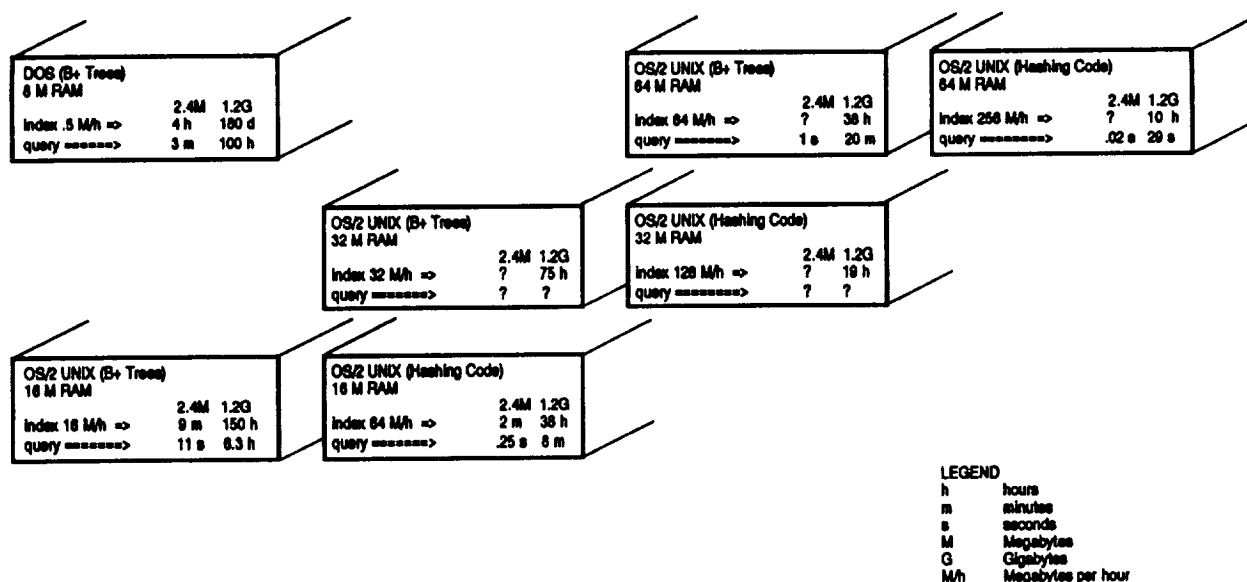


Figure 12. Run-Time Performance of the QA System.

References

- [1] C. Date, *An Introduction to Database Systems*, Vol. I, Addison Wesley, 1990.
- [2] Delphi Consulting Group, 1991, *Text Retrieval Systems: A Market and Technology Assessment*, 266 Beacon Street, Boston, MA, 1991.
- [3] J. Driscoll, J. Lautenschlager and M. Zhao, "The QA System," *Proc. of the First Text Retrieval Conference (TREC-1)*, NIST Special Publication 500-207 (D. K. Harman, editor), March, 1993.
- [4] D. Harman and G. Candela, "Retrieving Records from a Gigabyte of Text on a Minicomputer Using Statistical Ranking," *JASIS*, Vol. 41, pp. 581-589. 1990.
- [5] *Roget's International Thesaurus*, Harper & Row, New York, Fourth Edition, 1977.
- [6] G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill, 1968.
- [7] G. Salton, *The Smart Retrieval System—Experiments in Automatic Document Processing*, 1971.
- [8] G. Salton, E. A. Fox, and E. Voorhees, "Advanced Feedback Methods in Information Retrieval," *JASIS*, Vol. 36, pp. 200-210, 1985.
- [9] G. Salton, *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1989.
- [10] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," *JASIS*, Vol. 41, pp. 288-297, 1990.
- [11] *SPIRIT Version 2.1 User's Manual*, SYSTEX Company, Ferme Du Moulon, 91190 Gif Sur Yvette, France (French Edition), May 1986.
- [12] D. Voss and J. Driscoll, "Text Retrieval Using a Comprehensive Semantic Lexicon," *Proceedings of ISMM First International Conference on Information and Knowledge Management (CIKM-92)*, Baltimore, MD, November 1992.
- [13] E. Wendlandt and J. Driscoll, "Incorporating a Semantic Analysis into a Document Retrieval Strategy," *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Chicago, IL, pp. 270-279, October 1991.

DATABASE TOMOGRAPHY FOR COMMERCIAL APPLICATION

Dr. Ronald N. Kostoff
Director, Technical Assessment
Office Of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Mr. Henry J. Eberhart
Aerospace Engineer
Naval Air Warfare Center Weapons Division, Code C02422
1 Administration Circle
China Lake, CA 93555-6001

The views expressed in this paper are solely those of the authors and do not represent the views of the Department of the Navy.

ABSTRACT

Database Tomography is a revolutionary method for extracting themes and their relationships from text. The algorithms employed begin with word frequency and word proximity analysis and build upon these results. When the word 'database' is used, think of medical or police records, patents, journals, or papers, etc. (any text information that can be computer stored). Database Tomography features a full text, user interactive technique enabling the user to identify areas of interest, establish relationships, and map trends for a deeper understanding of an area of interest. Database Tomography concepts and applications have been reported in journals and presented at conferences. One important feature of the Database Tomography algorithm is that it can be used on a database of any size, and will facilitate the users ability to understand the volume of content therein. While employing the process to identify research opportunities it became obvious that this promising technology has potential applications for business, science, engineering, law, and academe. Examples include evaluating marketing trends, strategies, relationships and associations. Also, the Database Tomography process would be a powerful component in the areas of competitive intelligence, national security intelligence and patent analysis. User interests and involvement cannot be overemphasized.

INTRODUCTION

The purpose of this paper is to introduce the Database Tomography process to the commercial community and to other departments of the government. Database Tomography is a revolutionary method for extracting themes and their relationships from text. The algorithms employed start with word frequency and word proximity analysis and build upon these results. This process was developed under the sponsorship of the Office of Naval Research for the purpose of exploring trends in naval research and also identifying promising research opportunities. It was while employing the process for research opportunities that it became obvious that this promising technology has potential applications to the commercial sector and to other government departments. For example, it could be used for evaluating marketing trends, strategies, relationships and associations. In addition, we believe that the Database Tomography process would be a powerful component in the areas of competitive intelligence, national security intelligence, and patent analysis.

Database Tomography is a full text, user interactive technique that enables the user to identify areas of interest, establish relationships, and map trends so as to provide the user with a deeper understanding of his area of interest. The reader is referred to the bibliography section of this paper for examples of Database Tomography concepts and applications that have been reported in journals and presented at conferences. User interests shape the way in which the process is used, and this user involvement cannot be overemphasized.

Generalized and discipline-specific computer databases have become a major resource for researchers and analysts in business, science, engineering, law, and academe. Database size can make analysis and interpretation difficult. One important feature of the Database Tomography algorithm is that it can be used on a database of any size and can also facilitate the user's ability to understand the volume of content therein. When the word 'database' is used, the reader is encouraged to think of journals, papers, memos, reports, medical or police records, and patents—in other words, any text information that can be stored on a computer.

An application to a Former Soviet Union (FSU) text database is shown. This text describes a broad spectrum of FSU science (35 reports generated by the Foreign Applied Sciences Assessment Center (FASAC)). The algorithm extracts words and word phrases which are repeated throughout this large database. It allows the user to create a taxonomy of pervasive research thrusts from this extracted data. The algorithm then extracts words and phrases which occur **physically close** to the pervasive research thrusts throughout the text. It allows the user to determine interconnectivity among research thrusts, as well as determine research sub-thrusts strongly related to the pervasive thrusts.

The focus of the present study was to identify **technical thrusts** and their interrelationships. The raw data obtained by the extraction algorithms allowed the user to relate **technical thrusts to institutions, journals, people, geographical locations, and other categories**. The methodology can be applied to any text database, consisting of published papers, reports, and memos.

Background

About a decade ago, the U.S. Federal Government established the Foreign Applied Sciences Assessment Center (FASAC) under the operation of the Science Applications International Corporation (SAIC). The purpose of FASAC was to increase awareness of new foreign technologies with military, economic, or political importance. The emphasis was placed on "exploratory research" (Department of Defense 6.1/6.2 equivalent) in the FSU. This work seeks to translate fundamental research into new technology.

One of the main products of FASAC is reports on different areas of "exploratory research." FASAC assembles panels of expert consultants from academia, industry, and government. Each panel provides a written assessment of the status and potential impacts of foreign applied science in selected areas. Periodically, an Integration Report is generated that describes the trends in foreign research, including pervasive issues which affect research capabilities. By early 1992, there were about 40 reports on different aspects of FSU applied science.

CO-WORD ANALYSIS

Co-word analysis utilizes the proximity of words and their frequency of co-occurrence in some domain (sentence, paragraph, paper) to estimate the strength of their relationship. When applied to the literature in a technical field, co-word analysis allows a map of the relationship among technical themes to be constructed. A history of co-word analysis applied to research policy issues and co-word analysis origins in computational linguistics will be published in a Special Issue of *Competitive Intelligence Review* on technology. Over the past year, a **full text** word association technique (database tomography, a variant of co-word analysis) has been developed by the authors to allow rapid scanning of large text databases. The initial purpose of this development was to identify pervasive research thrusts (thrusts which transcend disciplines) from those large text databases which contain descriptions of many research programs or areas of research. Two applications have been reported:

1. Identification of pervasive research thrusts in a database describing promising research opportunities for the Navy. The database consisted of thirty reports produced by the National Academy of Sciences panels and Office of Naval Research (ONR) internal experts on 15 technical disciplines.
2. Identification of pervasive thrusts in the 7400 project Industrial R&D (IR&D) database. Applications to other large databases of (mainly) research program descriptions are ongoing.

The reported studies and the present study have used the following procedure:

First, the frequencies of appearance in the total text of all single words (for example, MATRIX), adjacent double words (METAL MATRIX), and adjacent triple words (METAL MATRIX COMPOSITES) are computed. The highest frequency technical content words are selected as the pervasive themes of the full database (for example, SHOCK WAVE, REMOTE SENSING, IMAGE PROCESSING).

Second, for each theme word, the frequencies of words within ± 50 words of the theme word for every occurrence in the full text are computed. A word frequency dictionary is constructed which shows the words closely related to the theme word. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each dictionary (hereafter called cluster) yield those subthemes closely related to the main cluster theme.

Third, threshold values are assigned to the numerical indices. These indices are used to filter out the most closely related words to the cluster theme (e.g., see Figure 1 for part of a typical filtered cluster from the FASAC study).

Cij	CI	II	Eij	CLUSTER MEMBER
		Cij/CI	$Cij^2/CICj$	
022	0036	0.611	0.0359	THERMAL INFRARED
056	0323	0.173	0.0259	ICE
070	0522	0.134	0.0250	SATELLITE

CODE:

Cij IS CO-OCCURRENCE FREQUENCY, OR NUMBER OF TIMES CLUSTER MEMBER APPEARS WITHIN ± 50 WORDS OF CLUSTER THEME IN TOTAL TEXT;

CI IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER MEMBER;

Cj IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER THEME;

II, THE INCLUSION INDEX BASED ON CLUSTER MEMBER, IS RATIO OF Cij TO CI; AND

Eij, THE EQUIVALENCE INDEX, IS PRODUCT OF INCLUSION INDEX BASED ON CLUSTER MEMBER II (Cij/CI) AND INCLUSION INDEX BASED ON CLUSTER THEME Ij (Cij/Cj).

Figure 1. Remote sensing cluster – closely related words.

Subsets of closely related words are combined into one file. Words which are common to more than one subset (cluster overlaps) are identified. Megaclusters, or strings of overlapping clusters (based on a threshold of numbers of common words, or overlaps), are constructed. These show umbrella areas of related research.

The final results identify:

1. The pervasive themes of the database
2. The relationship among these themes
3. The relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes.

Numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes. The qualitative analyses of the extracted data have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allows an understanding of the theme interrelationships not heretofore possible with previous text abstraction techniques using index or key words.

Application of Co-word Analysis to FASAC Database

The FSU is a major contributor to many areas of science and technology. FASAC reports help to document and provide insight to these contributions. There is present interest in preserving the basic science capability of the FSU. This task would benefit from improved understanding of the FSU science and technology capability.

Application of full text co-word analysis to the FSU component of the FASAC database could provide a unique perspective on the FSU science and technology capability. This database has a different structure from the databases analyzed previously. FASAC contains topical area assessments, whereas, the other databases analyzed contain program, project, or promising opportunity descriptions. Full text co-word analysis is sufficiently powerful and flexible to be applicable to FASAC as well. (Unclassified FASAC reports were used.) The FASAC database has a moderate density of technical terms. Most are scientific, but there are many institute names, journal names, publishers, and people names. Determination of the relationship among only technical areas is more difficult than in some purely technically focused databases which were analyzed previously. However, the data allows analyses which go beyond purely technical relationships.

MULTIWORD FREQUENCY ANALYSIS

The output of the multiword frequency analysis allows construction of a multilevel taxonomy of the full database. This taxonomy derives from the language and natural divisions of the database (analogous to a natural coordinate system of the database). Database entries are easily categorized. Other taxonomies are generated top-down and usually attempt to force-fit database subjects into pre-determined categories.

One advantage of the present full text approach over the index or key word approach is that many types of taxonomies can be generated, such as:

- Science
- Technology
- Institution
- Journal
- Person name

Within any one of these categories, such as science, many types of taxonomies can be developed. An example of one science taxonomy of the FASAC database will be shown.

Based on the high frequency single, adjacent double and triple words, the following high level taxonomy was generated. The capitalized words are *sample* high frequency words from the multiword frequency analyses:

- Information:
 - DATA
 - IMAGE PROCESSING
 - STATISTICAL PATTERN RECOGNITION
- Physics:
 - LASER
 - SHOCK WAVE
 - CHARGED PARTICLE ACCELERATORS
- Environment:
 - OCEAN
 - SEA SURFACE
 - INTERNAL GRAVITY WAVES

- **Materials:**
 - **MATERIALS**
 - **THIN FILM**
 - **METAL MATRIX COMPOSITES**

Caution must be exercised in relating the above taxonomy based on FASAC to the actual taxonomy of all of FSU science. The FASAC reports represent selected areas of FSU science. We do not know how representative all the FASAC reports are of total FSU science. The FASAC reports tend to reflect the open FSU literature. We do not know how well this open literature represents all of FSU science, including classified work and other unreported work.

The above taxonomy reflects frequency of word usage. It represents the numbers of words written about technical areas in the FASAC reports. Dollars spent on these areas, or other measures of FSU priorities, were not taken into account. The taxonomy could be skewed relative to FSU importance attached to these areas. Nevertheless, the above taxonomy does offer insight into areas of FSU science of interest to the U.S.

Megaclusters

Clusters which had three or more overlaps (three or more common members) were combined to form strings of related clusters, or megaclusters. The following megaclusters were obtained:

- Ionospheric Heating/Modification
- Image/Optical Processing
- Air-Sea Interface
- Low Observable
- Explosive Combustion
- Particle Beams
- Automatic/Remote Control
- Frequency Standards
- Radar Cross Section

Of the 60 cluster themes that were used to compute overlaps, 52 were in one of the nine megaclusters above. Most of the eight remaining themes could be subsumed under the nine megaclusters.

The science discipline taxonomy for the FASAC database was derived from the multiword frequency analysis. It was defined as Information, Physics, Environment, and Materials. In terms of the megaclusters:

- Information would encompass:
 - IMAGE/OPTICAL PROCESSING
 - AUTOMATIC/REMOTE CONTROL
- Physics would encompass:
 - IONOSPHERIC HEATING/MODIFICATION
 - PARTICLE BEAMS
 - FREQUENCY STANDARDS
 - RADAR CROSS SECTION
- Environment would encompass:
 - AIR-SEA INTERFACE
- Materials would encompass:
 - EXPLOSIVE COMBUSTION
 - LOW OBSERVABLE

Categorizing the database with the megacluster subcategories allows a re-interpretation of the FASAC database. FASAC is a compendium of those aspects of FSU science of interest to the U.S. for strategic and military purposes rather than a microcosm of all of FSU science.

For example, many classes of materials were researched and developed in the FSU. Yet the materials subcategory in the FASAC analysis focuses on FSU capabilities in energetic materials (explosives and propellants) and coatings to reduce radar cross sections. Both classes are important from a military viewpoint. The main environmental focus is air-sea interface. There is little mention of the terrestrial environment. The primary information category focus is on image and optical processing, and the secondary information category focus is on remote control. We could conclude that the FASAC concern was FSU capability in sensing the ocean for ship and submarine activity, and remotely processing and interpreting this information.

The secondary environmental focus of FASAC was on the ionosphere. Specifically, it was on FSU capabilities for modifying the ionosphere through high power radio wave heating and exploiting its use as a communication medium. One focus of the physics category was particle beams. These could have dual applications of high energy directed weapons and heaters for magnetically confined plasmas and inertial fusion targets.

Cluster Theme/Member Relationships

The final display, Figure 2, shows high technical content words from one of the smallest of the 60 clusters. The selection cutoff criterion was an Equivalence Index (see Figure 1 for definition) greater than or equal to 0.001. A simple division of word categories into quadrants based on Inclusion Index values was used to display the relationships of the cluster members to the cluster theme and to each other.

In Figure 2, the underlined topic, ATMOS OCEANIC PHYS, is the cluster theme. The cluster members are segregated into quadrants headed by their values of Inclusion Indices. I_j is the ratio of C_{ij} to C_j , and is the Inclusion Index based on the theme word. I_i is the ratio of C_{ij} to C_i , and is the Inclusion Index based on the cluster member. The dividing points between high and low I_j and I_i are the middle of the "knee" of the distribution functions of numbers of cluster members vs. values of I_j and I_i . All cluster members with I_j greater than or equal to 0.1 were defined as having high I_j . All cluster members with I_i greater than or equal to 0.5 were defined as having high I_i .

<u>ATMOS OCEANIC PHYS</u> CLUSTER - HIGH TECHNICAL CONTENT WORDS					
HIGH I_j			HIGH I_i		
HIGH I_j	LOW I_i		LOW I_j	HIGH I_i	
SEA			RADIOACOUSTIC SOUNDING		
INTERNAL WAVE			ACOUSTIC SOUNDING		
ACOUSTIC			THEORY OF WIND		
SCATTERING			MODELING OF SURFACE		
RADAR			WIND WAVES ATMOS		
SEA SURFACE			INFRASOUND AND INTERNAL ATTENUATION OF SOUND		
ATMOSPHERE			THEORY OF WAVE		
LOW I_j			LOW I_i		
WIND WAVES			SHEAR FLOW		PROCESSING OF RADAR
SOUND PROPAGATION			TURBULENT		WAVE PROPAGATION
OCEAN SURFACE			SATELLITE		WIND VELOCITY
GRAVITY WAVES			INTERNAL GRAVITY WAVES		POINT SOURCE
STRATIFIED FLUID			SOUND WAVES		

Figure 2. High technical content words of final display.

A high value of I_j means that, whenever the theme word appears in the text, there is a high probability that the cluster member will appear within ± 50 words of the theme word. A high value of I_i means that, whenever the cluster member appears in the text, there is a high probability that the theme word will appear within ± 50 words of the cluster member.

Thus, words located in the upper quadrant (high I_j high I_i) are coupled very strongly to the theme word. Whenever the theme word appears, there is a high probability that the cluster member will be physically close. Whenever the cluster member appears, there is a high probability that the theme word will be physically close. Whenever either word appears in the text, the other will be physically close.

Consider words located in the left quadrant (high I_j low I_i). Whenever the cluster member appears in the text, there is a low probability that it will be physically close to the theme word. Whenever the theme word appears in the text, there is a high probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member C_i is substantially larger than the frequency of occurrence of the theme word C_j , and the cluster member and the theme word have some related meaning.

Single words have absolute frequencies of an order of magnitude higher than double words. Thus, the words in the left quadrant are typically high frequency single words. They are related to the theme word but much broader in meaning than the theme word. A small fraction of the time that these broad single words appear, the more narrowly defined double word theme will appear physically close. However, whenever the narrowly defined double word theme appears, the broader related single word cluster member will appear. The words in the left quadrant can also be viewed as a higher level taxonomy of technical disciplines related to the theme ATMOS OCEANIC PHYS.

Consider words located in the right quadrant (low I_j high I_i). Whenever the cluster member appears in the text, there is a high probability that it will be physically close to the theme word. Whenever the theme word appears in the text, there is a low probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member C_i is substantially smaller than the frequency of occurrence of the theme word C_j , and the cluster member and the theme word have some related meaning. Thus, the words in the right quadrant tend to be low frequency double and triple words, related to the theme word but very narrowly defined.

A large fraction of the time that these very narrow double and triple words appear, the relatively broader double word theme will appear physically close. However, a small fraction of the time that the relatively broad double word theme appears, the more narrow double and triple word cluster member will appear. This quadrant grouping has the potential for identifying "needle-in-a-haystack" type thrusts which occur infrequently but strongly support the theme when they do occur. One of many advantages of full text over key or index words is this illustrated ability to retain low frequency but highly important words, since the key word approach ignores the low frequency words.

The words in the bottom quadrant (low I_j low I_i) are the remainder of the culled words. They relate to and support the theme, but do not have the strong inclusions based on theme or cluster member occurrence of the members of the other quadrants. The upper quadrant typically contains very few or no words. The left quadrant contains very broad words related to the theme. The right quadrant contains extremely narrow words related to the theme. The bottom quadrant contains words related to the theme of the same level of specificity as the theme (on average).

Figure 2, ATMOS OCEANIC PHYS, has a null upper quadrant (typical of the majority of clusters for the threshold values of Equivalence index chosen). The left quadrant, the broad taxonomy of related areas, appears to describe two major thrusts:

1. Underwater related (SEA, INTERNAL WAVE, ACOUSTIC, SCATTERING) focusing on sound propagation through the sea.
2. Atmosphere related (ATMOSPHERE, RADAR, SEA SURFACE, SCATTERING) focusing on radar propagation through the atmosphere.

The thrusts have a common juncture at the sea surface, where both acoustic and radar scattering occur on different sides.

The right quadrant focuses on very specific subareas related primarily to acoustics. These include acoustics applied to the atmosphere (RADIOACOUSTIC SOUNDING), and other aspects of atmospheric science (THEORY OF WIND).

The bottom quadrant provides the most balanced view of the two thrusts. It expands on the underwater propagation medium (STRATIFIED FLUID, SHEAR FLOW, INTERNAL GRAVITY WAVES), the radar platform issues (SATELLITE, PROCESSING OF RADAR), and the ocean surface issues (WIND WAVES, TURBULENT, OCEAN SURFACE). The integrated picture presented by the three quadrants is the use of radar from a space platform to view the ocean surface, and the research problems arising from the wind and undersea flows governing the conditions and structure of the ocean surface and impacting the interpretation of the radar images.

CONCLUSIONS

Based on the results and interpretation of the multiword frequency analysis and the co-word analysis, the FASAC database used in this study is a compendium of those aspects of FSU science of interest to the U.S. for strategic and military purposes. The microlevel analysis of selected theme clusters, showing how the cluster members related to each theme, reinforced this conclusion and provided more detail about those aspects of each theme on which FASAC concentrated.

A wealth of information resulted from the FASAC output, and only a small fraction of that information was presented and analyzed in this paper. The analysis was restricted to technical themes and their relationships. Raw data was available for relating technical themes to non-technical themes such as institutions, scientists, journals, and geographical regions.

In the future, full text co-word analysis could be used to obtain a more representative structure of FSU (or any other country's) science. If a large number of randomly selected published FSU scientific papers were entered into a database, then a multiword frequency analysis and co-word analysis could be performed on this text database.

Assume that a paper represents about \$100K worth of effort. A 10,000 paper database would represent \$1B worth of effort, and would offer a very representative sample of FSU science output. The 10,000 paper database could be analyzed on an existing advanced desktop computer. The critical path would be assembling this database, not analyzing it.

Full text co-word analysis is in its formative stages. Much development remains to be done to understand the breadth of analyses which can be performed and the breadth of applications which can be covered. It is hoped that the initial techniques and results reported in this study will motivate and stimulate other organizations and researchers to develop and apply the general technique of full text co-word analysis on a much broader scale.

ACKNOWLEDGMENT

The authors wish to note the contribution of Mr. David Miles in creating the computer-based algorithm, *TEXT SLICER*.

BIBLIOGRAPHY

Kostoff, R. N. "Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis," Proceedings: *Portland International Conference on Management of Engineering and Technology*, 27-31 October 1991. More detailed paper available from author.

----- "Research Impact Assessment," Presented at *Third International Conference on Management of Technology*, Miami, FL, 17-21 February 1992. Larger text available from author.

Kostoff, R. N. "Co-Word Analysis for the Evaluation of R&D," in *Assessing R&D Impacts: Method and Practice*, B. Bozeman and J. Melkers, eds. Kluwer Academic Publishers, Norwell, MA, 1993a.

----- . "Database Tomography: Origins and Applications," *Competitive Intelligence Review*, Special Issue on Technology, 1993b, to be published.

----- . "Database Tomography for Technical Intelligence," *Competitive Intelligence Review*, Volume 4, Number 1, Spring 1993c.

----- . "Database Tomography: Origins And Applications," *Competitive Intelligence Review*, Special Issue on Technology, Volume 5, Number 1, Spring 1994, Wiley & Sons.

NAS. "Reorientation of the Research Capability of the Former Soviet Union," Results of a Workshop on 3 March 1992, NAS, NAE, IOM, National Academy Press, Washington, D.C., 1992.

AUTHOR IDENTIFICATION

Ronald Neil Kostoff received a Ph.D. in Aerospace and Mechanical Sciences from Princeton University in 1967. At Bell Labs, he performed technical studies in support of the Office of Manned Space Flight, and economic and financial studies in support of AT&T Headquarters. At the U.S. Department of Energy, he managed the Nuclear Applied Technology Development Division, the Fusion Systems Studies Program, and the Advanced Technology Program. At the Office of Naval Research, he is Director of Technical Assessment, and his present interests revolve around improved methods to assess the impact of research.

Henry Joseph Eberhart holds a Masters degree in Mechanical Engineering from the University of Southern California and a Bachelors degree in Engineering from the University of California, Los Angeles. His entire career has been with the Navy at China Lake, California and has involved materials, stress analysis, and test and evaluation of air-to-surface missile systems. He is presently engaged with the planning functions at what is now the Naval Air Warfare Center Weapons Division China Lake.

**AUTOMATED MAINFRAME DATA COLLECTION
IN A NETWORK ENVIRONMENT**

**David L. Gross
Computer Engineer
Analex Space Systems, Inc.
NASA Kennedy Space Center
Post Office Box 21206
Kennedy Space Center, FL 32815-0206
Phone: 407/861-5716
Fax: 407/861-5774**

ABSTRACT

The progress and direction of the computer industry have resulted in widespread use of dissimilar and incompatible mainframe data systems. Data collection from these multiple systems is a labor intensive task. In the past, data collection has been restricted to the efforts of personnel specially trained on each system. Information is one of the most important resources an organization has. Any improvement in an organization's ability to access and manage that information provides a competitive advantage. This problem of data collection is compounded at NASA sites by multi-center and contractor operations. The Centralized Automated Data Retrieval System (CADRS) is designed to provide a common interface that would permit data access, query, and retrieval from multiple contractor and NASA systems. The methods developed for CADRS have a strong commercial potential in that they would be applicable for any industry that needs inter-department, inter-company, or inter-agency data communications. The widespread use of multi-system data networks, that combine older legacy systems with newer decentralized networks, has made data retrieval a critical problem for information dependent industries. Implementing the technology discussed in this paper would reduce operational expenses and improve data collection on these composite data systems.

INTRODUCTION

The need to access and retrieve data from mainframe systems is a widespread labor intensive activity. A number of commercial products based on the client/server concept are available to solve this problem. In a client/server system the "client" portion of the applications reside on workstations or Local Area Networks (LAN) with the "server" portion running on larger machines (i.e. mainframes). Economically the cost of purchasing, installing, and maintaining such products on one or more systems can outweigh the savings in manhours. These systems do save time in data retrieval and system access but they require a significant initial investment in additional training, equipment, and software development tools. The cost and time required for data retrievals increase geometrically when multiple, usually dissimilar systems are integrated. Tying different systems together means connecting incompatible architectures, protocols and languages. This paper discusses a composite system that can perform many of the same retrieval functions of a client/server system but without the technical restrictions and financial overhead involved.

BACKGROUND

In 1990 NASA funded a project to improve the data retrieval and dissemination methods used by the Safety, Reliability & Quality Assurance (SR&QA) Directorate. The methods currently being used require highly skilled data researchers to access and query over 27 different mainframe systems. Data requests could take from a few minutes to a few days to complete depending on the number and types of systems involved. The goal of the project was to develop a more time-efficient method for performing these retrievals. Several commercial client/server systems were evaluated, but from a technical or cost perspective, they were unacceptable. The decision was made to develop a new method of data retrieval.

The Centralized Automated Data Retrieval System (CADRS) is a result of this project. CADRS is a network based system that automatically handles all system accesses, queries, data conversions, and transmittals from the mainframe to the PC environment. This implementation required the development of three sub-systems. These sub-systems are the Central Document Database (CDD), Automated Reporting System (ARS), and Forms Query (FQ).

The three subsystems work in tandem to fulfill all of the data handling requirements of the SRM&QA group. The Central Document Database acts as the primary user interface to all general data reports and supporting technical documentation. The updating of this information as well as single user event driven reports are handled by the Automated Reporting System (ARS). The Forms Query system performs all ad hoc (one time only) data searches and report generations.

CENTRAL DOCUMENT DATABASE

The CDD is a repository for all technical documents and reports that require easy access and full search capability. It provides the advanced document handling techniques and specialized search techniques required by the user community. This is a dynamic system that contains a library of technical documents and host data reports that can be accessed through a Local Area Network (see Figure 1). The system is designed to access documents and delimited data files of various sizes and types that are stored at different physical locations and provides a single interface for viewing and searching of this data.

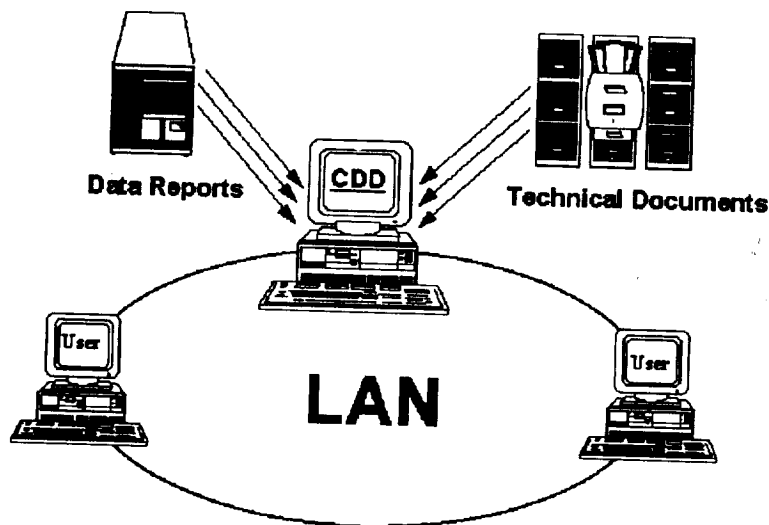


Figure 1 Data Flow of the Central Document Database

A basic menu system is used to call up and display all of the available documents and reports. The system uses a number of filters that enable it to access documents in common word processors and mainframe printer formats. These are simple filters designed to mask the specialized command codes used by the different application programs that produced the document. As a result a document in almost any format, (e.g. Word Perfect, Displaywrite, or mainframe redirected printer output), can be accessed by the system and displayed to the user.

The CDD has the capability to perform several different types and levels of data searches. The simplest search is a basic Boolean key-word search. This type of search is a useful and fairly common type of search that can locate a specific string using standard AND/OR logic. The program provides an improvement to this type of search by expanding the Boolean logic to include any acronyms and abbreviations of the user's search request from its built-in knowledge database.

The most complex search the program performs is based on a natural language parsing and weighting network. This network can identify and rank the key areas of the document that are most likely to contain the requested information. The program syntactically breaks down the user's data request and converts it to a network of related search words (see Figure 2). The document(s) being searched are then compared word by word to this network. A weight value is assigned to each word or phrase in the network. The sum of the weights for each word found in a section determines the overall value for that section. A list of pointers, to the sections of the document that had the highest values, is the final result of the filtering. The software will immediately display the area of the document that had the highest weighting. If the user does not find the desired information, he can move to the next highest weighted area.

Technical documents accessed through the CDD are initially stored by direct scanning using optical character recognition software or by downloading an existing online version. Most online documents are "dynamic" in that they are constantly being updated and revised. To ensure that the most up-to-date version of a document is available, the CDD interfaces with and uses the services of, the Automated Reporting System (ARS). The ARS automatically transmits the latest revisions of the documents to the CDD. This ensures that the CDD has the latest version of all stored technical documents as well as the latest data reports.

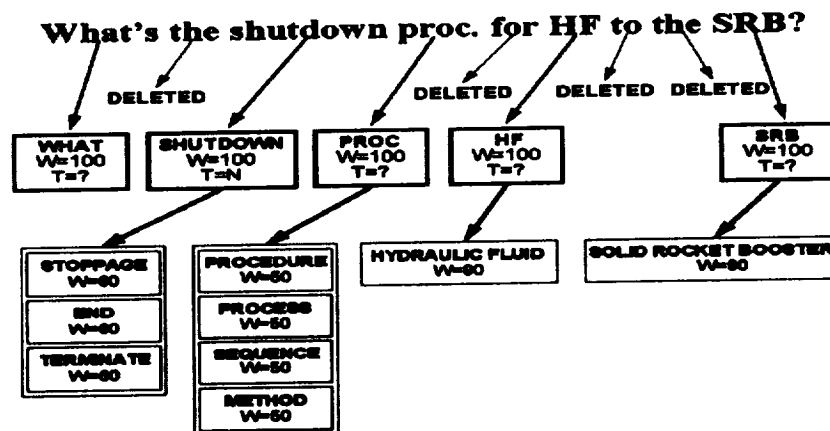


Figure 2 Development of the Weighting Network

AUTOMATED REPORTING SYSTEM

The ARS is an autonomous mainframe based set of tools and techniques designed to automatically query and transmit time or event driven data requests. The system runs on several major NASA host systems. It uses a collection of programs, queries, and scripts to collect data and transmit it to the user's Local Area Network (LAN). The host programs act as "initiators" and run as scheduled batch operations. These programs are automatically initiated on a daily, weekly, or "per shuttle flow" basis during non-peak periods of the day. This minimizes the system's impact on overtaxed mainframe resources. Below is an example of an "initiator" program written in the REXX (Restructured Extended Executor) interpreter language.

```
'SQLINIT D(KSC1H0P3'
'CP LINK SYSP01 19A RR'      /* PRODUCTION COMMON DISK */
'ACC 19A F'
'CP LINK SQLDBA 195 RR'      /* SQL/DS */
'ACC 195 G'
'CP LINK ISPVM 192 193 RR'    /* ISPF & ISPF/PDF */
'ACC 193 H'
'CP LINK MAINT 19D RR'       /* HELP */
'ACC 19D K'
'CP LINK MAINT 303 RR'       /* GDDM - GDDM/PGF - GDDM/GKS */
'CP LINK MAINT 303 RR'       /* GDDM - GDDM/PGF - GDDM/GKS */
'ACC 303 L'
'CP LINK SYSADMIN 399 RR'    /* PROFS */
'ACC 399 M'
'CP LINK MAINT 347 RR'       /* QMF */
'ACC 347 N'
THISDATE = DATE('S')
W = DATE('B')//7
Y = SUBSTR(THISDATE,1,4)
M = SUBSTR(THISDATE,5,2)
D = SUBSTR(THISDATE,7,2)
NDAY = '31 28 31 30 31 30 31 30 31 30 31'
IF W = 0 THEN,                /* CURRENT DAY IS MONDAY */
    COUNT = 3
ELSE
    COUNT = 1
DO A = 1 TO COUNT
    D = D - 1
    IF D = 0 THEN DO
        M = M - 1
        IF M = 0 THEN DO
            M = '12'
            Y = Y - 1
        END
        D = WORD(NDAY,M)
    END
END
YESTERDAY = Y||'-'||M||'-'||D
SAY YESTERDAY
'REPORTER AR005 (US <CTRID>|<CTRID> DISK AR005 TEXT A PARML ' YESTERDAY
'SF AR005 TEXT A TO CDD1 AT RQVMKSC'
```

This program is designed to be invoked by VMSchedul running on a IBM 3090 every weekday before first shift. The program performs all required links, identifies the previous weekday date and passes the date as a parameter to a DBreporter query. The data returned from the query are transmitted to the CDD address on the local network.

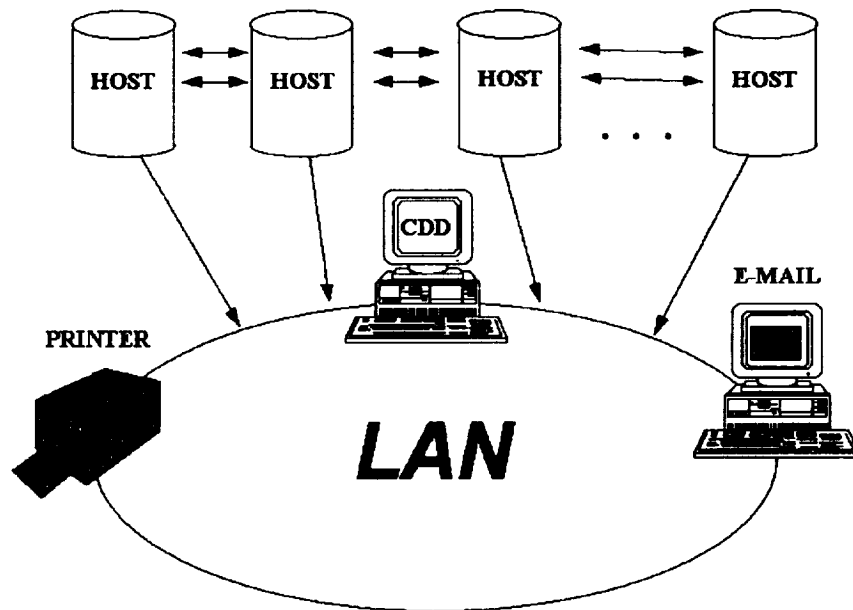


Figure 3 Data Transfer Paths of the ARS

Procedures are initiated and run against host database systems in a host-compatible language (SQL, LOUIS, Dbreporter). The resulting data is formatted into reports and transmitted through electronic routing paths to the user's local e_mail address, office printers, and the CDD (see Figure 3). The user's requirements determines the actual location where the data is to be transmitted. The normal policy is to have data reports intended for single users transmitted via e-mail or sent to a local printer. All other reports are sent into the CDD for general access.

FORMS QUERY

The Forms Query (FQ) system is a user interface linked through the CADRS server to several host "open" queries. Open queries are standardized common data queries with the actual parameter values missing. The user interface generates a file of these missing values which it transmits to the host system (see Figure 4). The basic user interface has been built around the concept of "forms." This is the use of a graphical representation of existing NASA forms for querying and displaying data (see Figure 5). All existing "Form" screens access a legal value dictionary and field identification system. This provides a context sensitive help and data identification system for

the display. This also serves to restrict requests to actual legal values and data ranges. The interface program and host communication procedures were developed using C, Pascal, and Enfin development software. These access procedures support both Token Ring and synchronous communication lines.

The screenshot shows a graphical user interface window titled "PROBLEM REPORT". The window has a menu bar with "File", "Control", "Clear", and "Submit". Below the menu bar, there are logos for NASA and USAF, and the text "Kennedy Space Center/Andrews Air Force Base". The form contains several input fields and checkboxes:

- PROBLEM REPORT**: A label above a text field containing "PV-8-80078".
- DETECTED DURING**: A text field.
- WORK AREA**: A text field containing "PAD-A".
- END ITEM CONTROL NUMBER**: A text field containing "APU-20-5631".
- WORK UNIT CODE**: A text field.
- PART/PROG NAME**: A text field.
- PART/PROG NO**: A text field.
- SER/REV NO**: A text field.
- QTY**: A text field.
- FSCM/VENDOR**: A text field.
- NHAFN/TAPE/DISC**: A text field.
- STS #EFF**: A text field containing "30 ± 105 ±".
- REPORTED BY**: A text field.
- DATE**: A text field.
- SOFTWARE PROBLEM LOCATOR**: A section with checkboxes for "DUMP", "TRANSLATOR OUT.", "LINE PRINTER OUT.", "COMPILER LISTING", and "OTHER".
- PROBLEM DESCRIPTION**: A large text area for describing the problem.
- CRIT SKILLS**: A checkbox.
- CHANGE REQ**: A checkbox.
- CONSTRAINTS**: A checkbox.
- YES**: A checkbox.
- CRIT**: A checkbox.
- RESP ORG.**: A text field.

Figure 4 Example Form in the Forms Query System

Once the user has entered all his data requirements into the form it is submitted to the CADRS server. The server converts the data into a set of variable values for a related open query. These values, along with the query template identification and user identification are transmitted to the indicated host system. An "initiator" program on the host system checks for this information on a periodic basis. When the program finds a parameter file waiting it submits the indicated open query with the data passed as parameters. The results of the query are transmitted down to the CADRS server where they can be accessed by the user.

The multi-tasking ability of Windows 3.1 allows the user to continue with other tasks while the query is being processed. When the data has been received the user can access it using the same forms the initial query request was made from. A number of additional utilities have been included in the system. These include a limited charting capability and data export facilities. Built-in data conversion functions from the ENFIN software libraries have been incorporated into the Query Forms system menus. This allows direct data conversion between common mainframe host data formats and workstation systems. Currently mainframe data can be converted into dbase, RBase, Oracle, Excel, SQL Server, and delimited ASCII. This allows the data to be imported into popular graphing and charting programs.

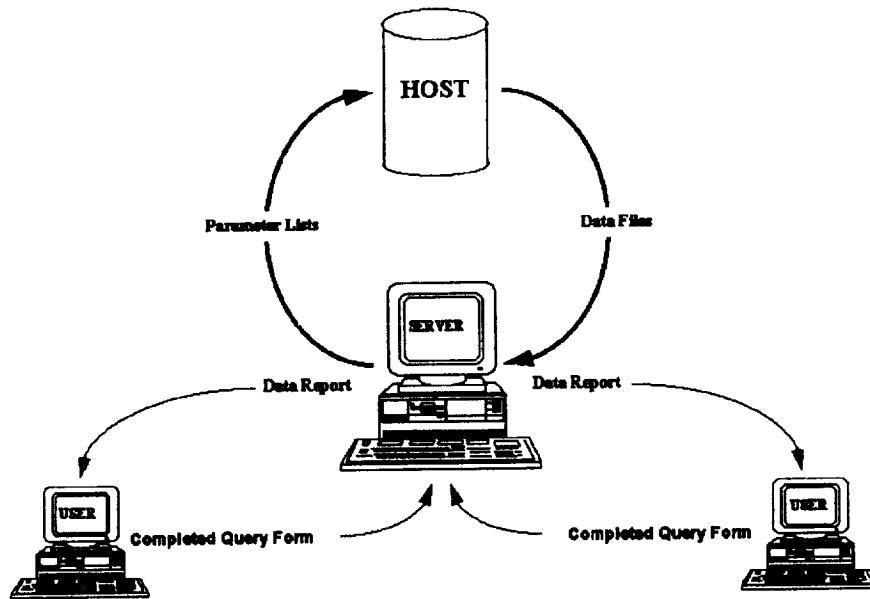


Figure 5 Data Interface of the Forms Query System

IMPLEMENTATION

The Central Document Database is installed on a Compaq SystemPro server. A one Gigabyte optical disk drive is used for local data and document storage. Communication links to mainframe systems are made through a Token Ring network and asynchronous data lines. The CDD server is connected to user workstations through Ethernet and Microsoft's Lan Manager software.

The Automated Reporting System activates itself every 24 hours during a non-peak time periods. Initiation is performed by using existing mainframe scheduling products (i.e. VMSchedul in IBM's VM system). A log command file is processed that identifies queries, data format procedures, and electronic addresses. On the local area network the CDD intercepts all data transmitted from ARS to its address. The CDD incorporates these data reports into its own database. Transmittals to individual users and network printers are identified by their own network address.

The Forms Query system resides both on the workstation as a user interface application and as open queries on the different mainframe systems. An executable procedure residing on each of the host systems checks at frequent time periods (one hour) for any parameter files sent from the interface application. These files identify the query, user and all variable values needed. The queries are run with the indicated variable values from the parameter file. Any data that are returned from the query is transmitted back to the local area network tagged with the user's ID.

SUMMARY

The CADRS system provides a unique solution to the problems of dissimilar and incompatible host systems, was compounded at NASA sites by multi-center and contractor operations. Currently, there are 27 different mainframe systems in widespread use by the space program. Included in this number are NASA specific systems as well as in-house contractor systems. Although, the situation at NASA is unusual it is by no means unique. Commercial industry with multiple legacy systems would find CADRS to be a viable option for data retrieval and dissemination. The system provides a low cost alternative to client/server systems when information retrieval is the primary consideration.

The three sub-systems of CADRS can be operated as a stand-alone system to provide improved data access. The CDD can be used on stand alone workstations to handle technical documents and manuals. Its ability to perform intelligent searches on large documents makes it well suited for reference systems. The ARS system provides techniques to automate standard data retrieval processes. This provides man-hour savings as well as shifting resource intensive tasks to non-peak periods. The Forms Query system provides a low cost graphical interface for performing common queries. These forms allow non-trained personnel to perform a greater percentage of the required data retrievals. Whether using all or a part of CADRS the benefits of the technology are obvious.

REFERENCES

- [1] Gross, D., "Methods and Means Used in Programming Intelligent Searches of Technical Documents", NASA Conference Publication 3189, Volume 2, pp 55-62, 1992.
- [2] Butler, J., "Middleware Needed to Plug Client/Server Holes", Software Magazine, July 1993 v13 n10 pg 55-57.
- [3] Luger, G. and Stubblefield, W., "Artificial Intelligence and the Design of Expert Systems", Benjamin/Cummings Publishing Company, Inc. 1989.
- [4] Minsky, M., "Semantic Information Processing", The Massachusetts Institute of Technology Press, 1968.
- [5] Whittington, R.P., "Database Systems Engineering", Oxford University Press, 1988.

**BEGINNING THE 21ST CENTURY
WITH ADVANCED AUTOMATIC PARTS IDENTIFICATION (API)**

**Fred Schramm
Technology Utilization Office
Marshall Space Flight Center, AL 35812**

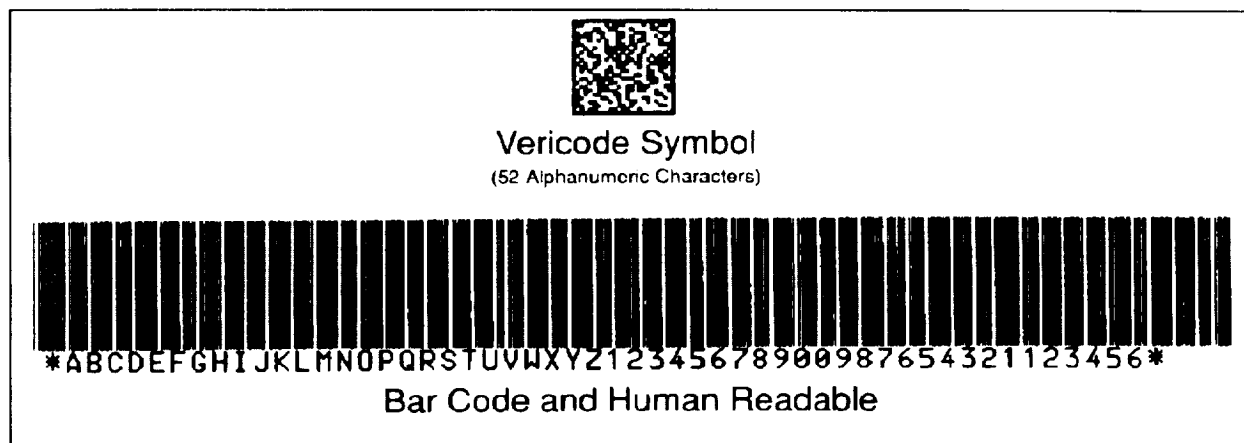
**Don Roxby
Rockwell International
555 Discovery Drive
Huntsville, AL 35806**

ABSTRACT

Under the direction of the NASA George C. Marshall Space Flight Center, Huntsville, Alabama, the development and commercialization of an advanced Automated Parts Identification (API) system is being undertaken by Rockwell International Corporation. The new API system is based on a variable sized, machine-readable, two-dimensional matrix symbol that can be applied directly onto most metallic and nonmetallic materials using safe, permanent marking methods. Its checkerboard-like structure is the most space efficient of all symbologies. This high data-density symbology can be applied to products of different material sizes and geometries using application-dependent, computer-driven marking devices. The high fidelity markings produced by these devices can then be captured using a specially designed camera linked to any IBM-compatible computer. Application of compressed symbology technology will reduce costs and improve quality, productivity, and processes in a wide variety of federal and commercial applications.

Existing Automated Identification Systems

There are thousands of applications for automatic identification. Although many technologies are available, most currently use bar code systems. Bar codes are one-dimensional systems and are generally attached to products using paper labels or tags, or by incorporating the code onto the product wrapper. This indirect marking approach, while suitable for retail sales, distribution, and other applications that are not paperless, is inadequate for marking products subject to harsh environments and handling. Bar coded paper labels, for example, are not tolerant of heat, cold, rain, wind, abrasion, chemicals, and other unfriendly conditions many products encounter during their life cycles. In addition to the limitations of typical bar code label material, the basic bar code design—long code length, fixed size and lack of error correction—has its own set of limitations when the decoding system attempts to deal with a variety of substrates.



A comparison of VERICODE® and bar code symbologies using the same 52-character string.

Advanced Automated Parts Identification Systems

Because of inherent limitations of bar code for direct part marking in the aerospace industry, NASA sought a more suitable API technology. In 1991, Rockwell was asked by NASA to initiate a compressed symbology testing program to survey the API industry, evaluate the commercially available products, and choose an approach that would satisfy component marking requirements in the Space Shuttle program. The test program was open to all manufacturers of advanced API systems that met the specific criteria developed for NASA aerospace program applications. These criteria included, but were not limited to, the following:

1. The system software must be capable of producing user-selectable data densities to meet a wide-range of applications, such as six to 30 character-codes.
2. The symbol structure must be conducive to the application of symbols onto metallic and nonmetallic substrates exhibiting a wide range of surface reflectivity and topography, using the permanent hardware marking methods defined by MIL-STD-130 and others.
3. The system must be capable of producing readable, low data-density symbols small enough to apply directly onto electronic components and critical fasteners.
4. The system software must provide symbol damage reconstruction features to overcome the various mechanisms of symbol deterioration that can be expected in aerospace applications, e.g., fading, scratches, gouges, etc.
5. The system must be capable of reading from acute angles of camera and symbol rotation under a variety of environmental conditions.
6. The system must provide the ability to read symbols in remote locations requiring camera portability, e.g., internal vehicle structures, stock yards, launch pads, emergency landing strips, etc.

In August 1991, the newly created Rockwell Huntsville Compressed Symbology Laboratory conducted the initial pretest system screening. Nine different systems were reviewed; all but two were eliminated. The selected systems were two-dimensional (matrix) systems and were subsequently brought on-site for evaluation. The test program that followed resulted in the VeriSystem® being chosen as best suited for an advanced API system for aerospace applications.

VERICODE® System (VeriSystem®) Description

The VERICODE® Symbol, developed by Veritec Inc., Chatsworth, California, is a high data-density, two-dimensional, machine-readable code that can be produced in variable size. The symbol can be applied directly to virtually any material including metal, plastics, glass, paper, etc. and read using a fixed station or portable charged coupled device (CCD) camera. Designed to run on IBM-compatible microcomputers, the VeriSystem® consists of two basic modules: encode, which converts the human-readable code into VERICODE®; and decode, which reverses the translation process.

The encode process converts human-readable data, e.g., serial or inventory tracking number, into a VERICODE® Symbol. The user may specify the data density and the size of the symbol required to fit the available marking space. After the symbol is prepared by the encoding process, it is marked on a part using a marking methodology appropriate for the part's composition and anticipated operating environment. The VERICODE® Symbol can be permanently affixed to virtually any material regardless of shape.

The decode process captures the symbol via CCD camera and then recognizes and converts it. The VeriSystem® can "read" symbols under a wide variety of orientations. Symbols can be rotated up to 360 degrees and tilted up to 80 degrees from the horizontal position. Error detection and correction capabilities of VeriSystem® will salvage as much of the data as possible and identify missing or damaged data. The system also

features data reconstruction that allows for recognition, regeneration, and decoding of damaged symbols. The reconstruction feature allows for recovery up to 40 percent by adding data codes.

Decoded data can be transferred to a host computer via hard wire, radio frequency (RF) or modem. If the decoded data is a component's serial number, the unique number can be used as a key to retrieve life cycle, maintenance, or other information about that particular component from a data base. This decoded data can also be used by computerized inventory, manufacturing, and schedules' subsystems in an organization.

PROGRAM APPLICATIONS AND BENEFITS

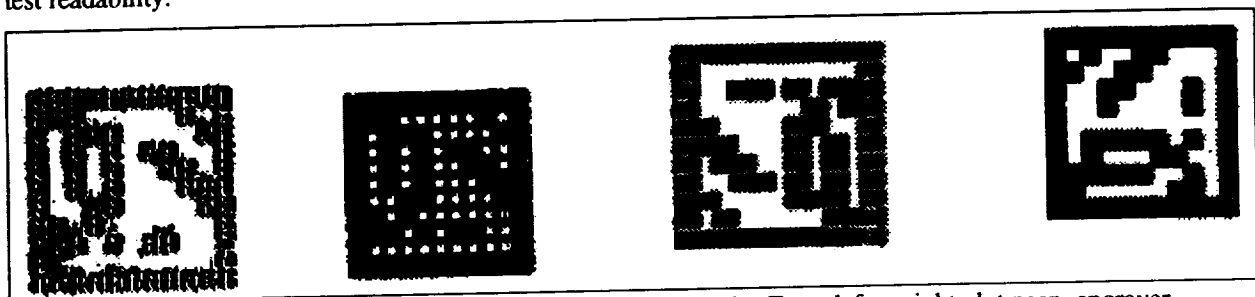
The majority of aerospace technologies are product specific and were primarily targeted at solving specific problems using exotic materials, unique processes, and numerous government specifications. However, compressed symbology is a dual-use technology, meaning it has a wide range of applications in both federal and commercial industries. The technology's ability to automate parts identification can improve competitiveness and provide a quick return on investment.

Compressed symbology provides a computer-based foundation for a true closed-loop, fully automated configuration management and accounting system. The VeriSystem® eliminates the need for labels and tags on small components, piece parts and materials; and eliminates manufacturing operations' need for concurrent paper travel, such as work authorization documents that travel with items during fabrication, assembly, checkout, installation, and changeout. The product, and the information about the product, flows together—synchronized. The code logic is virtually tamper proof.

Compressed symbology technology is ideally suited to inventory marking of parts and components in the aerospace, automotive, electronics, and pharmaceutical industries. Incorporation of compressed symbology in computer-based manufacturing, fabrication, and assembly applications can eliminate a major source of errors—data reentry—by eliminating the manual reentry of component identification data. Compressed symbology can add significant value to automated design and manufacturing, configuration management, modular tooling, robotics, and quality control.

DIRECT PART MARKING—DEVELOPMENT STATUS

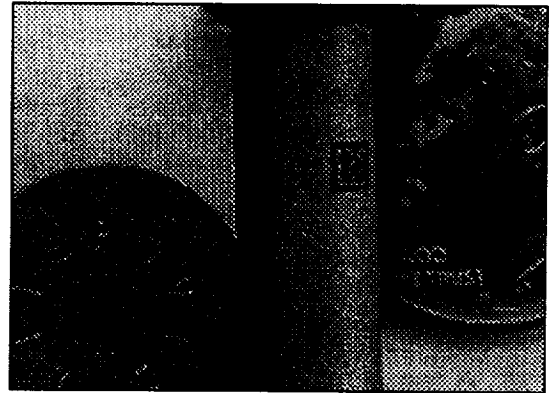
Direct part marking has the potential of damaging the substrate, a risk not encountered with indirect marking approaches. The goal of the marking process is to make a readable, durable mark (in essence, a controlled defect) without disturbing the surrounding substrate. A variety of materials, most of which are routinely used in the Space Shuttle program, have been subjected to extensive metallurgy, fatigue, and environmental testing to assess the affects of different marking methods. These tests are used to determine the marking methods and marking device settings that produce the best marks with the least change to the material's surface. Fatigue tests are performed to ensure that the marks do not significantly change material properties and therefore reduce the life of the part. Environmental tests assess the readability of the marks subjected to environmental conditions expected during the life cycle of the material. For example, environmental tests, consisting primarily of long-term exposure to salt spray, are being performed for the Defense Industrial Supply Center on marked fasteners used by the military to test readability.



A bar of Aluminum 2024 showing four direct part marking methods. From left to right, dot peen, engraver (backfilled with flat black paint), micro-abrasive, and Nd:YAG laser.

Thirty-three marking methods for applying VERICODE® Symbols were identified, but only six computer-based approaches were selected for testing. The six are:

1. **Laser Etch**—A highly reliable technology, the laser etch marking method has been in use since 1976. The AB Laser Company Landmark Model 6001 computer-driven Nd:YAG laser marking system can produce quality markings at precise depths by adjusting power, frequency, and speed settings. These adjustments provide the control necessary to apply machine-readable VERICODE® Symbols, human-readable markings, and graphics to most surfaces with a minimum of surface disruption. With current technology, the smallest readable symbol produced by the laser etch marking method is 0.125-inch square.



VERICODE® Symbols etched by laser on critical fasteners.

2. **Machine Engraving**—The Newing-Hall, Inc. model NH-300 computer-driven engraving system is used to remove base material to a prescribed depth and is accomplished using a computer-guided cutting tool. The method is suitable for marking VERICODE® Symbols, human-readable markings, and graphics onto most metallic and nonmetallic substrates used on the Space Shuttle program. Readability of the symbols is contingent upon the contrast between the surface material and engraving. In instances where there is no natural contrast, a backfill material (permanent ink, paint, etc.) is applied to the engraved mark. With current technology, the smallest readable symbol produced by the machine-engraving marking method is 0.125-inch square.

3. **Dot Peen**—The Wetzel Tool Company Mark V computer-driven dot peen marking system produces machine readable VERICODE® Symbols, human-readable markings, and graphics (logos) on metallic and nonmetallic materials. The system utilizes a computer-operated device that drives a pointed tungsten carbide stylus onto the surface to be marked. The device marks the material by repeatedly striking the surface with the stylus, forming a series of dots (cone shaped recesses). The highly localized, low stress blows deform and stretch the metal surface causing a difference in contrast between the peened and unpeened areas of the part. This difference in contrast can normally be read and decoded by the VeriSystem®. Dot peened materials, which do not naturally provide adequate contrast for decoding, can be enhanced by backfilling the recessed areas with a permanent filler material of contrasting color. With current technology, the smallest readable symbol produced by the dot peen marking method is 0.125-inch square.



VERICODE® Symbol dot peened onto a turbine blade.

4. **Micro-Abrasive**—The Comco, Inc. model PR1101-3 computer-driven micro-abrasive system is a miniature “sand” blaster that marks by directing a controlled stream of abrasive over the surface of the material to be marked. Dependent upon the abrasive used, and the settings made, the system can cut or texture VERICODE® Symbols and human-readable markings onto metallic or nonmetallic surfaces. The system operates by directing a mixture of dry air and abrasive material through a small tungsten carbide nozzle at high velocity. Software automatically controls the movement of the

marking nozzle, the length of the stop and go pulses, and the speed of flow to produce the requested marking. The system can be adjusted to accurately control the depth of cut. With current technology, the smallest readable symbol produced by the micro-abrasive marking method is 0.375-inch square.

5. **Fabric Embroidery**—The fabric embroidery marking method involves the use of a computer-driven sewing machine to stitch a representation of the VERICODE® Symbol onto cloth or fabric materials. The marking technique is a simple, cost-effective method for applying durable markings onto cloth and fabric materials. The fabric embroidery marking method could be extremely effective for marking Thermal Protection System (TPS) blankets. Rockwell currently identifies TPS blankets using the ink stamp marking method. The quality of the marking is generally poor due to the coarseness of the fabric and the markings do not hold up well over time. Efforts are now underway to locate heat resistant black thread that can be used in the fabric embroidery marking process. With current technology, the smallest readable symbol produced by the fabric embroidery marking method is 0.50-inch square.
6. **Ink Jet**—The Jet Equipment Corporation SMS-92-H Standard Marking System can produce high-resolution markings on a wide variety of topographies with no adverse effect on the substrate. This fast, non-contact marking method uses a computer-controlled X/Y/Z marking head to deposit industrial grade ink that dries almost immediately. The ink jet marking method has been approved for use by the aerospace industry. With current technology, the smallest readable symbol produced by the ink jet marking method is 0.125-inch square.



VERICODE® Symbol machine-embroidered on a Thermal Protection System blanket.

Special Coatings

In addition to laboratory tests, there are a number of tests being performed under aerospace production and flight conditions. For instance, there are eight VERICODE® marked TPS tiles being flown on three orbiters in the NASA Space Shuttle fleet (see Figure 1). The TPS tiles are used to protect the orbiters' aluminum skin at launch, in earth orbit, and during the fiery re-entries encountered before landing—one of the most hostile environments an identification mark can encounter. During normal orbiter flight cycles, TPS tiles are subjected to adverse weather conditions on the launch pads, temperature extremes that range from -250 to +2200 degrees Fahrenheit, air flows that can exceed 18,000 miles per hour, and abrasion from dirt and sand during the launch and recovery process. A special thermal coating, Verishield® TM/CTI 43B, designed to withstand extreme variations in temperature on a variety of materials, was developed by Ceram Tech International Ltd., Pomona, New York, to meet TPS requirements. So far, the VERICODE® Symbols have remained readable after multiple Space Shuttle missions.

Phased Implementation

Compressed symbology is being implemented on the turbopump manufacturing process at Rockwell International, Rocketdyne Division, Canoga Park, California. The project activities will include developing compressed symbology-based part identification standards, environmental testing of VERICODE® Symbols marked directly on critical parts and materials, and electronically transferring part information between a shop floor manufacturing environment and a remote computer data base. The Rockwell Huntsville Compressed Symbology Laboratory is testing a portable data capture and transmission system to be used in the field. The device is a small, rugged portable computer linked to a handheld scanner capable of transmitting data to remote host computers via cable or RF transmission (wireless). Also, the Rockwell Huntsville Compressed Symbology Laboratory and Rocketdyne are developing software that interfaces the commercial VeriSystem® to the SSME configuration management systems. Using the VeriSystem® with its direct parts marking capabilities, the automated identification, authentication, and traceability of a part in the manufacturing and assembly process, as well as throughout its life cycle, will be possible.



Space Shuttle Discovery

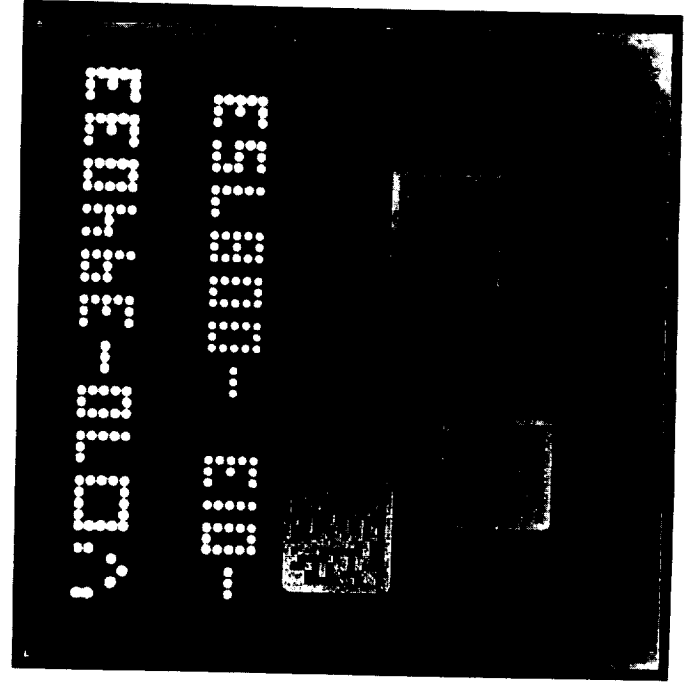


Figure 1: Thermal Protection System tiles marked with VERICODE® Symbols have remained readable after multiple Space Shuttle missions. The symbols were etched into the surface of the tile with a laser and backfilled with a thermal coating called Verishield®

Contingent on test results, the Marshall Space Flight Center Shuttle projects (External Tank, Solid Rocket Booster, and Solid Rocket Motor, etc.) plan the implementation of compressed symbology in current operations where feasible and on all future programs.

Outreach

Compressed Symbology is a dual-use technology that has a wide range of applications in both federal and commercial industries. Because the Rockwell Huntsville Compressed Symbology Laboratory is funded by the government, the technology and experience gained in the lab is available to the public. The Outreach program attempts to transfer technology, gained through tax dollars, into the public and private sectors. Any federal or commercial entity can request assistance on a specific problem by submitting a "Problem Statement" through the Marshall Space Flight Center Technology Utilization Office. Problem statements related to materials marking or compressed symbology are forwarded to the Rockwell Huntsville Compressed Symbology Laboratory where a viable solution is sought. The Rockwell Huntsville Compressed Symbology Laboratory has already responded to Technology Utilization problem statements from all branches of the military, government agencies like NASA, USDA, and FAA, universities, and private industry. In the last sixteen months, the Rockwell Huntsville Compressed Symbology Laboratory has been asked to mark a wide range of materials: from animal carcasses and Jack Daniel's wooden whisky barrels, to small electronics components (resistors, diodes, capacitors) and dental drills.

From ancient cave drawings until 1960, man has kept up with things manually. Yesterday's manual modus operandi was time-consuming, error-prone, and expensive. Today's approach, ushered in by the first practical applications of bar code in 1960, is characterized by bar codes, computers, and paper labels. Bar code is a giant step toward automating yesterday's manual processes, but its technical limitations make it unsuitable for many aerospace, automotive, electronics, and other industrial applications. Tomorrow's approach requires a more advanced API technology. The technology must support direct part marking, data recovery, a higher degree of computerization, and a wide range of "read" orientations and distances. Compressed symbology is the advanced API that NASA and Rockwell International will employ as they reach toward the year 2003 and beyond.

omit

PHOTONICS

OPTICAL PROCESSING FOR SEMICONDUCTOR DEVICE FABRICATION

Bhushan L. Sopori

National Renewable Energy Laboratory
1617 Cole Boulevard
Golden, CO 80401

ABSTRACT

A new technique for semiconductor device processing is described that uses optical energy to produce local heating/melting in the vicinity of a preselected interface of the device. This process, called Optical Processing, invokes assistance of photons to enhance interface reactions such as diffusion and melting, as compared to the use of thermal heating alone. Optical processing is performed in a "cold wall" furnace, and requires considerably lower energies than furnace or rapid thermal annealing. This technique can produce some device structures with unique properties that cannot be produced by conventional thermal processing. Some applications of Optical Processing involving semiconductor-metal interfaces are described.

INTRODUCTION

Fabrication of electronic devices requires process steps such as diffusion, oxidation, and contact formation. During these process steps the desired reactions take place only at the interfaces. For example, in a junction formation process the dopant diffusion may take place only a fraction of a micron deep; likewise, oxides are typically grown less than 1000Å thick (1-3). Even though the reactions are primarily limited to interfaces, the conventional approach for carrying out various device fabrication process steps involves heating the entire wafer in a furnace to the process temperature. A recent method, rapid thermal annealing (RTA), also heats the entire wafer nearly isothermally; however, due to the use of optical heating, the rate-of-increase in the temperature can be very rapid (4). Thus, in spite of the fact that the device fabrication often requires reaction(s) to take place only near the surface or interfaces, the methods developed to date require that the entire wafer be heated. Any attempts to locally heat the material by producing large steady-state temperature gradients can generate large stresses and, hence, produce defects in the material or even shatter the process wafers. The only somewhat successful approach for local heating consists of heating by means of very short pulses from a short wavelength laser. Even with laser heating, only the front surface can be locally heated. Due to many disadvantages such as low throughput, long coherence length of the laser light, and high cost of laser processing, this technology is still in the research mode.

This paper describes a new processing technique that can preferentially deliver energy to an interface of a semiconductor (S) and a metal (M), even if the interface is deep inside the material. The energy delivered to the interface can be controlled to modify the interface characteristics. The major emphasis of this paper is to describe some applications of this processing technique for formation of low-resistivity ohmic contacts on semiconductors with some unique properties.

OPTICAL PROCESSING

In Optical Processing, the S-M interface is illuminated from the semiconductor side with a spectrum such that the major part of the light reaches the interface. The incident light produces local heating, accompanied by enhanced diffusion and/or melting, in the vicinity of the interface. The thickness of the melt or the diffusion depth can be controlled by controlling the energy delivered to the device. The local melt can be generated to form an alloyed region that regrows epitaxially on the silicon substrate and produces an ohmic contact of extremely low contact resistivity. The energy delivered to the device can also produce bulk heating to induce other

predetermined thermal effects. The interface reaction is strongly diminished if the S-M interface is not directly illuminated.

Optical Processing and Rapid Thermal Annealing (RTA) differ in the basic mechanisms involved in each process. In Optical Processing the reaction at the interface is assisted by photons. Hence, the reaction occurs predominantly at the illuminated interface; the same reaction is greatly slowed if the interface is masked. In contrast, a typical RTA is a thermal process that cannot discriminate between the front and the backside of the device since such a process is completely thermally controlled [4].

Optical Processing can be best understood by an example of its application. Here we will consider an example of simultaneous formation of ohmic contacts to a silicon solar cell.

FABRICATION OF CONTACTS ON SOLAR CELLS

Fabrication of low-resistance metal contacts on solar cells requires sintering and alloying of the S-M interfaces to produce the desired ohmic characteristics. In the fabrication of a typical n+/p solar cell, the back metal must be alloyed to have a low-resistivity contact on the higher-resistivity base region, while the front contact must be only mildly sintered to prevent metal from punching through the highly doped emitter, and the depletion regions. The need for different processing conditions for the front and back contacts necessitates several process steps, the number depending on the method of metal deposition (e.g., plating, screen printing, or evaporation). As an example, Table 1 shows the typical steps involved in a conventional process using evaporated aluminum to make contacts on both sides. Also indicated in the table are the process steps for making the contacts by optical processing, as discussed below.

Table I: Comparison of process steps for fabricating contacts to solar cells using conventional methods and Optical Processing (this example uses Al on both sides).

<u>FURNACE ANNEAL OR RTA</u>	<u>OPTICAL PROCESSING</u>
1 Deposit Al on back side	Deposit Al on front and back
2 First alloy	Sinter/alloy
3 Strip excess Al in HCl then Rinse in DI water; dry	
4 Dil. HF dip (or fume)	
5 Deposit Al on front	
6 Sinter front Al	
7 Deposit additional Al on back	

Figure 1 illustrates the structure of an n+/p solar cell which has a thin layer of Al deposited on the entire backside and narrow Al pads on the front side. In order to produce high quality contacts to the cell, the back contact must be alloyed without alloying the front. To accomplish this, the cell is placed in an Optical Processing Furnace (OPF) with the junction side upward, as shown schematically in Figure 2. The OPF consists of a quartz muffle that is illuminated from above by quartz-halogen lamps. The optics of the light sources are designed so that the illumination in the process zone is highly uniform. Process gases such as Ar, N₂, and O₂ are regulated to flow through the furnace. The walls of the muffle are maintained "cold" by flowing N₂ along the outside walls of the muffle. The spectrum, intensity, and duration of the incident flux are chosen for the specific application. Figure 3 shows typical process cycles using Al contacts for silicon solar cells with and without antireflection coating. As seen from this figure, the process is controlled in terms of the optical power delivered to the device.

RESULTS

Figure 4 schematically illustrates the effects of Optical Processing cycles, described in Figure 3, on the front and back interfaces of a solar cell. While the front contact does not show the presence of an alloyed region, the back contact forms a thin Si-Al alloyed region adjacent to the Si surface. Figure 5 is a high resolution cross-sectional TEM image of the Si/Si-Al alloy interface showing an epitaxial growth of the alloy. By controlling the light intensity, and the process time, one can control the thickness of the alloyed back layer and the sinter conditions for the front contact simultaneously. In order to compare the degree of reaction at the front and the back side of the cell we have processed some devices under accentuated conditions. Figures 6 and 7 show the Al and Si profiles of these "strongly" processed contacts before and after removal of the residual Al. Figures 6a and 6b show that the back contact has melted to produce an alloyed interface region of about $0.4\text{ }\mu\text{m}$. However, Figures 7a and 7b show that the front contact has an alloyed region of $<< 0.1\text{ }\mu\text{m}$. It is important to point out that under normal process conditions the alloyed interface thicknesses are considerably smaller than those in Figures 6 and 7.

The quality of the contacts formed by Optical Processing is extremely high. From the initial measurements of contacts fabricated on $0.5\text{--}10\text{ }\Omega\text{-cm}$ substrates, we have estimated the contact resistivity to be less than $10^{-4}\text{ }\Omega\text{-cm}^2$, up to current densities of 2 A/cm^2 . As a result, solar cell contacts made by Optical Processing have excellent characteristics. Figure 8 shows the I-V characteristic of a large-area 32 cm^2 cell with contacts produced by this technique. The efficiency of the cell is slightly above that for the same cell with contacts made with conventional process steps. Optical Processing has also been applied to already fabricated contacts on fully finished commercial cells resulting in the improvements in the cell performance.

Although we have only discussed formation of Si-Al contacts, the same principle is applicable to many other semiconductor-metals combinations. We have used other metals, such as Cu, Ni, and Pd, in various combinations with silicon and obtained excellent results.

SPECIAL PROPERTIES OF METAL CONTACTS FORMED BY OPTICAL PROCESSING

The above described results involving controlled melt and diffusion at a semiconductor interface permits fabrication of some unique structures. Here we describe examples of two regimes of Optical Processing to control the S-M interface properties - the melt regime, and the diffusion regime. In the melt regime the interface is provided with sufficient energy to create a uniform melt that spreads laterally over the entire S-M interface. When the source of energy is removed, the melted region regrows epitaxially over the substrates. In the diffusion regime, the energy dissipated at the interface cannot produce a melt; however, localized regions of the enhanced diffusion across the interface can take place forming low-energy sites of nucleation. For example, enhanced diffusion at a Si-Al interface can create transport of Si to define etch pits bounded by (111) planes. These mechanisms can be used to produce following characteristics of the devices.

A. High reflectance ohmic contacts:

Typically low-resistivity ohmic contacts, produced by the conventional processing, result in graded interfaces that allow light to be transmitted from semiconductor into the metal where it can be absorbed. Optical processing allows the interface to be made abrupt by confining the melt to less than $100\text{ }\text{\AA}$. Such an interface can reflect light very effectively, and yet have an excellent ohmic characteristics with low resistivity. Figure 9 shows reflectivity of an optically processed contact as measured from the silicon side and from the aluminum side. The reflectivity from silicon side, at wavelengths larger than the $1.1\text{ }\mu\text{m}$, is about 80%;

corresponding reflectivity for a conventionally processed contact is typically less than 50%. This figure also shows that aluminum remains highly reflective after optical processing.

B. Dry texturing:

Optical Processing of a contact in the diffusion regime (followed by a second process step involving melt regime) produces dry texturing of the interface with a highly reflecting ohmic contact. Such a contact can efficiently scatter light to produce light trapping useful for thin-film solar cells. Figures 10A, 10B, and 10C are the photographs of the texture, produced under different process conditions, showing the control of the texture size and the density. Figure 10D is a higher magnification photograph showing pyramid shape of the texture formed on a (100) silicon wafer.

ADVANTAGES OF OPTICAL PROCESSING

Advantages of Optical Processing over conventional processes include

- Because heating/melting initiates at the interface (and can be confined to a thin region at the interface), the effect of the impurities in the ambient gas(es) on the characteristics of the contact is minimal compared to either furnace processing or Rapid Thermal Annealing (RTA). In Optical Processing the surfaces of the Al contacts typically remain shiny and do not require further preparation for additional metallization, such as solder dip (see Figure 9);
- Optical Processing is a "cold wall" process which minimizes the impurity out-diffusion as well as permeation from furnace walls;
- The process results in large-area uniformity of the alloyed/sintered layers. This feature is evidenced by the fact that the Si-Al contacts produced are free from the "spikes" and pitting produced by other processes;
- The process requires much less power than furnace or RTA anneals;
- The process is rapid, has high throughput, and can make devices with unique characteristics;
- The process requires fewer steps than conventional approaches and results in significant cost savings;
- Optical Processing allows control of dimensions both in depth and laterally. This is due, in part, to the fact that interactions can be induced to occur at considerably lower temperatures. This feature is important for VLSI and ULSI applications;
- This technique lends itself to multi-layer contact formation.

Optical Processing appears to have a strong commercial potential. This processing technique has already developed strong interest in the photovoltaic industry for commercial applications. This technology is also expected to have major applications in microelectronic device fabrication.

ACKNOWLEDGEMENT

The author would like to acknowledge many contributions to this work by Kim Jones, Sally Asher, Robert Reedy, and Doug Rose of NREL. This work was supported by NREL Technology Maturation Fund and by the U.S. Department of Energy under contract # DE-AC02-83CH10093.

REFERENCES

- [1] L. Sardi, S. Bargioni, C. Canali, P. Davoli, M. Prudenziati and V. Valbusa, *Solar Cells*, **11**, 51 (1984)
- [2] M. G. Coleman, R. A. Pryor, and T. G. Sparks, *Proc. 14th IEEE PVSC*, 793 (1980)
- [3] J. H. Wohlgemuth and S. Narayanan, *Proc. 22nd IEEE PVSC*, 273 (1991); and references therein.
- [4] For example, "Rapid Thermal and Integrated Processing," edited by J. C. Gelpey, M. L. Green, R. Singh, and J. J. Wortman, *Mat. Res. Soc. Symp. Proc.* **224**, (1991)

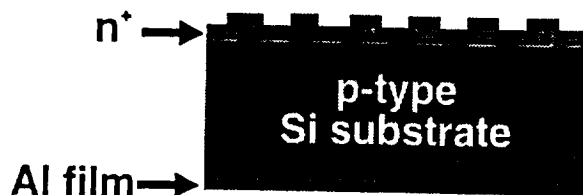


Figure 1. Schematic of the solar cell configuration used for simultaneous contact formation by Optical Processing

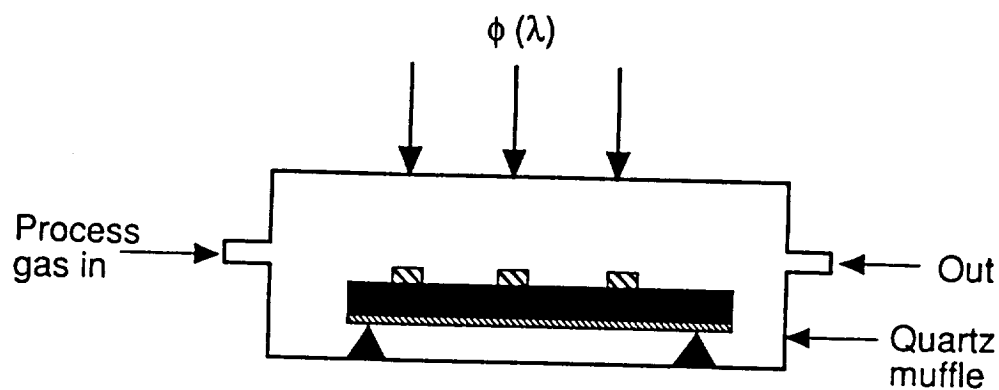


Figure 2. Illustration of the use Optical Processing for the simultaneous formation of front and back contacts

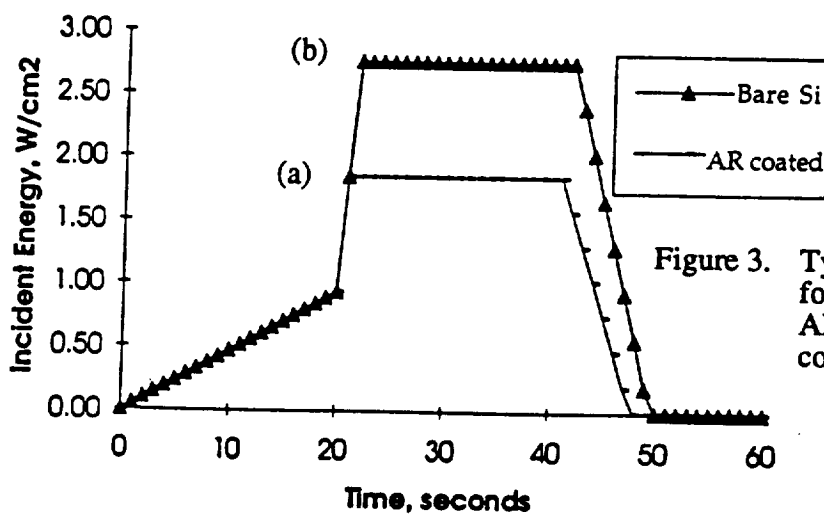


Figure 3. Typical process cycles for forming Si-Al contacts: (a) with AR coating; (b) without AR coating

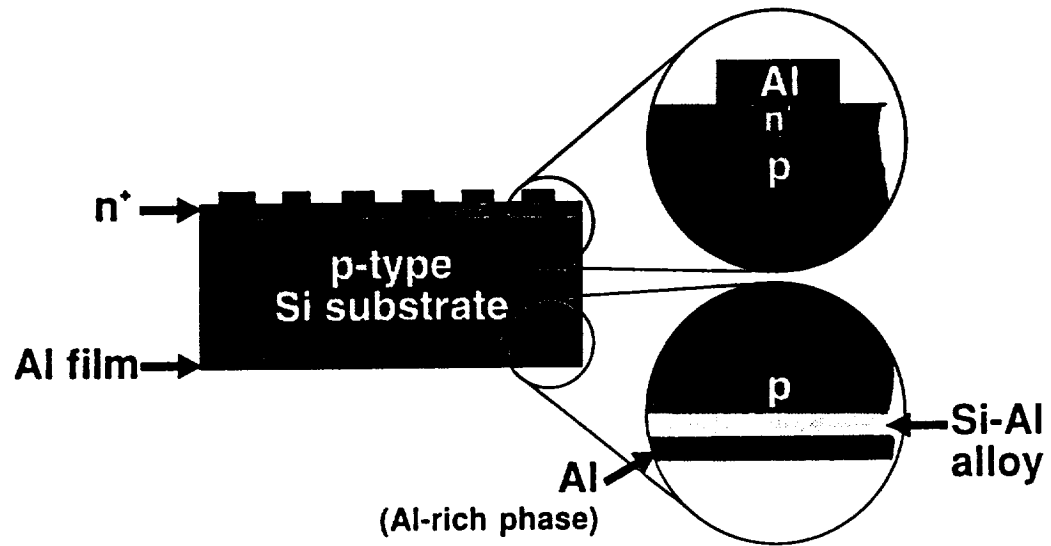


Figure 4. Illustration of the interface structure produced by Optical Processing

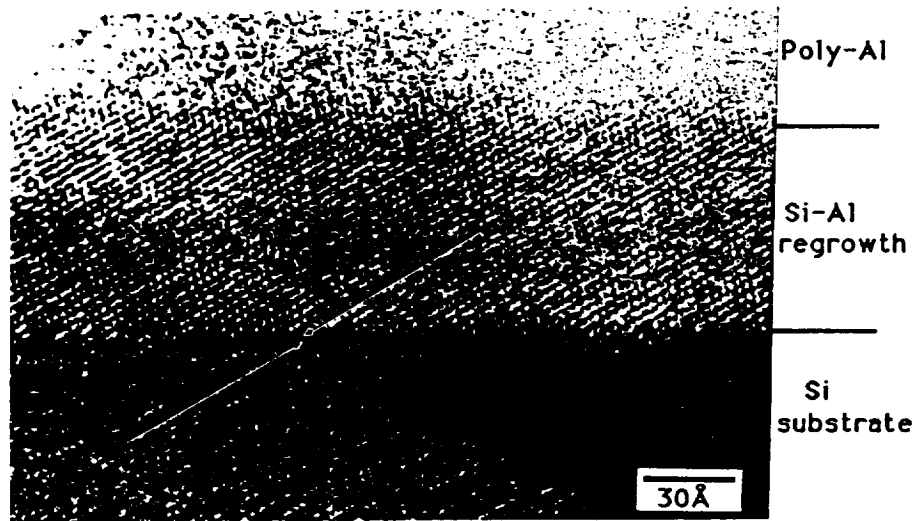


Figure 5. High resolution XTEM image of the Si/Si-Al alloy, showing epitaxial growth of the alloy on the substrate

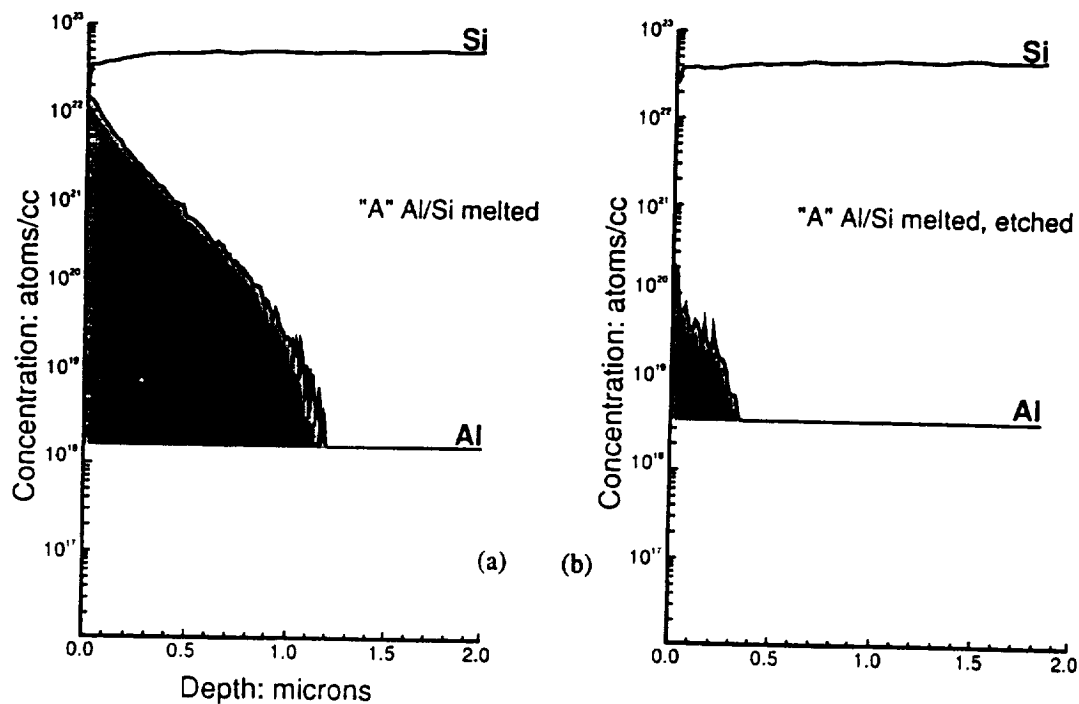


Figure 6. SIMS profiles of Al and Si on the backside contact of a "strongly" processed sample: (a) as-formed; (b) after residual Al was etched

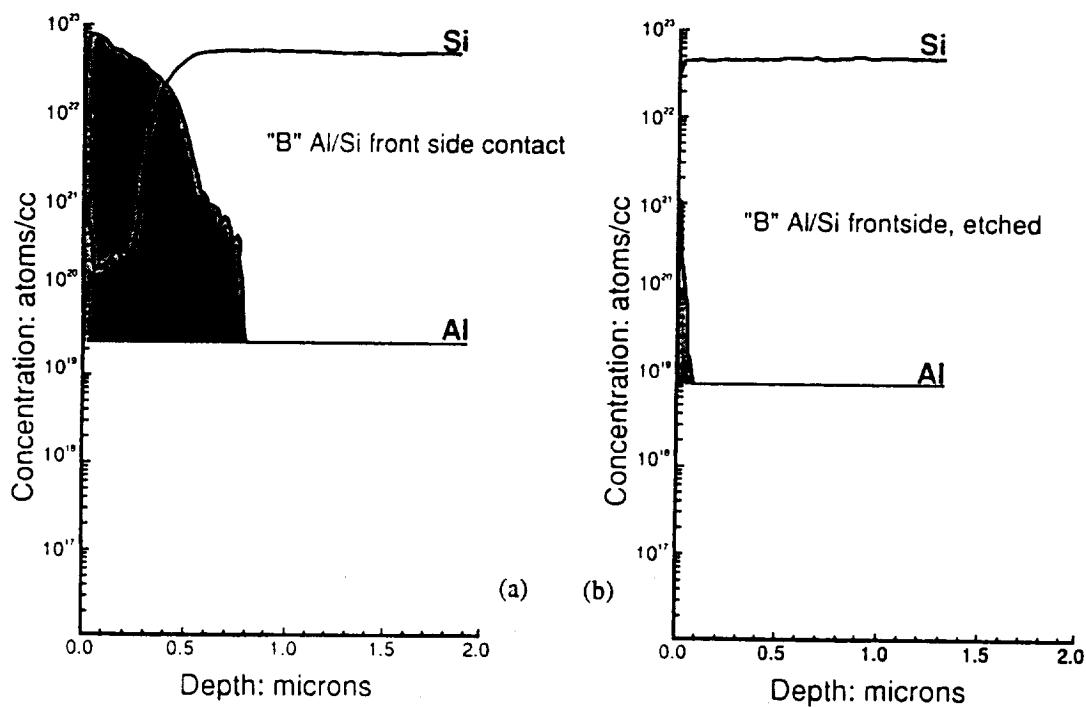


Figure 7. SIMS profiles of Al and Si on the front contact of a "strongly" processed sample: (a) as-formed; (b) after residual Al was etched

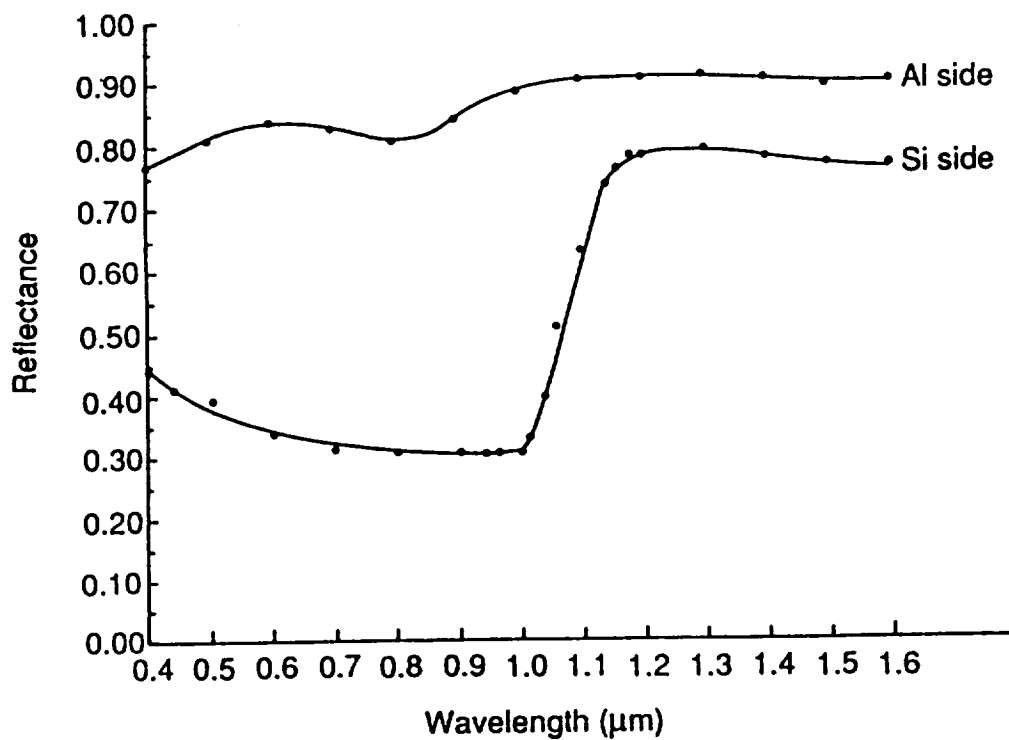
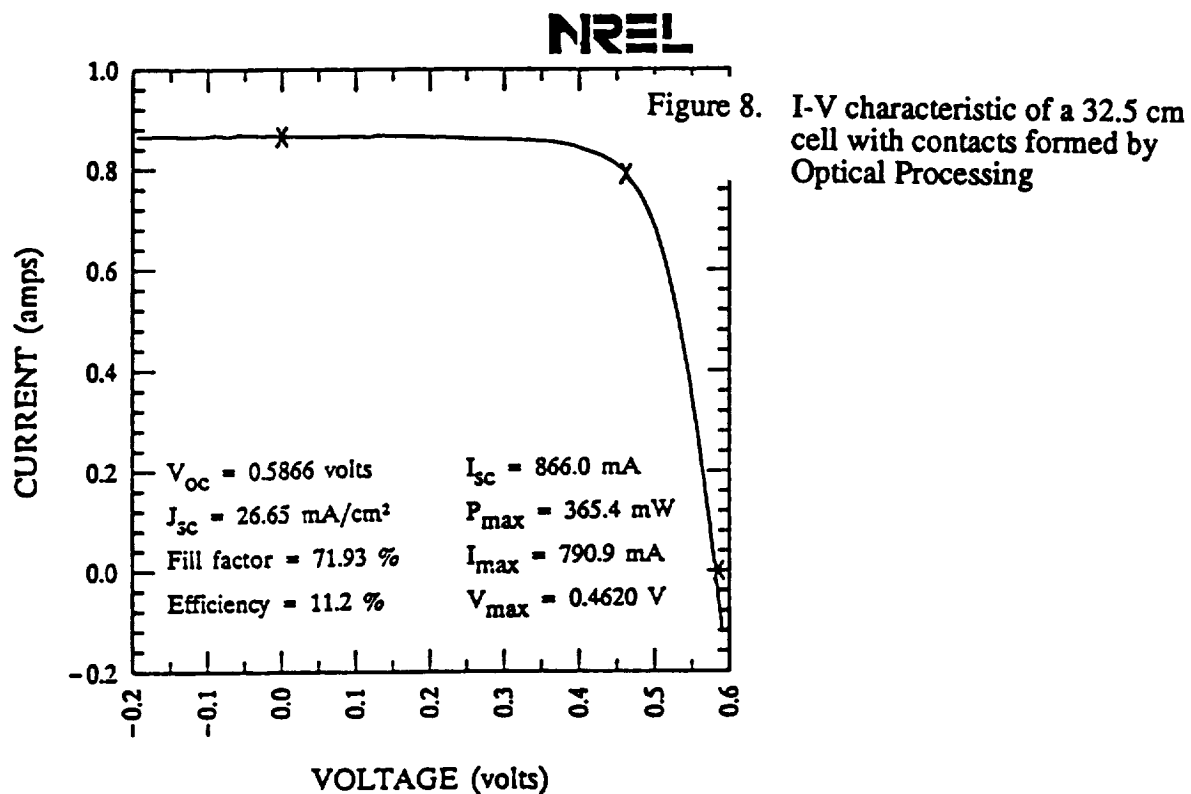


Figure 9. Reflectance plot of a high-reflectivity contact (a) from silicon side, (b) from Al side

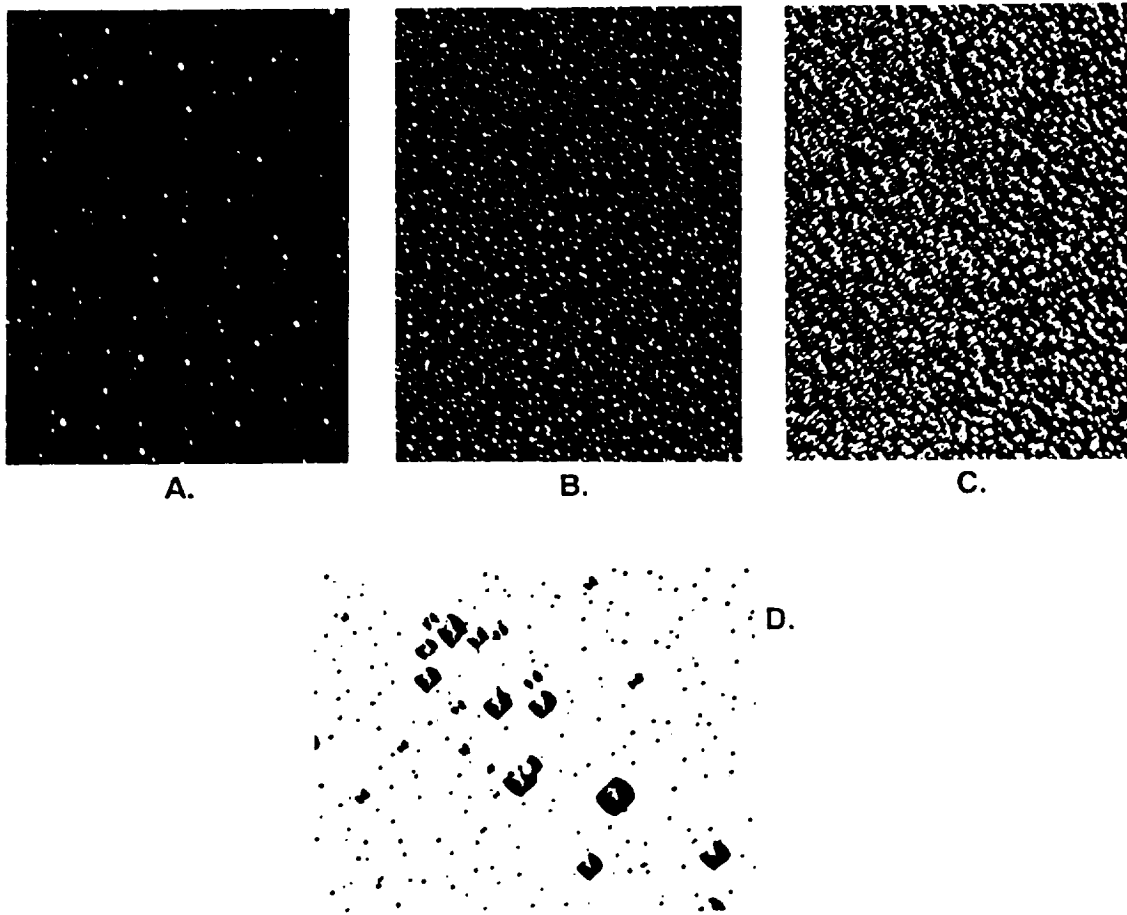


Figure 10. Photographs showing variation in the texture density and size due to different Optical Processing conditions (see text)

**A SCANNING DEFECT MAPPING SYSTEM
FOR SEMICONDUCTOR CHARACTERIZATION**

Bhushan L. Sopori

**National Renewable Energy Laboratory
1617 Cole Boulevard
Golden, Colorado 80401**

ABSTRACT

We have developed an optical scanning system that generates maps of the spatial distributions of defects in single and polycrystalline silicon wafers. This instrument, called Scanning Defect Mapping System, utilizes differences in the scattering characteristics of dislocation etch pits and grain boundaries from a defect-etched sample to identify, and count them. This system simultaneously operates in the dislocation mode and the grain boundary (GB) mode. In the "dislocation mode," the optical scattering from the etch pits is used to statistically count dislocations, while ignoring the GB's. Likewise, in the "grain boundary mode" the system only recognizes the local scattering from the GB's to generate grain boundary distributions. The information generated by this instrument is valuable for material quality control, identifying mechanisms of defect generation and the nature of thermal stresses during the crystal growth, and the solar cell process design.

INTRODUCTION

Crystal defects such as dislocations and grain boundaries strongly influence the performance of all electronic devices. The influence of defects is particularly important for solar cells because the commercial silicon solar cells are fabricated on low-quality material that contains high densities of defects and impurities. In the large-grain polycrystalline silicon substrates, used for commercial solar cells, the dominant defect appears to be the intragrain dislocations. A knowledge of the distribution of defects is necessary for the following reasons: (i) the distribution of defects reflects the nature of thermal stresses generated during the crystal growth. It is known that dislocation formation can take place when the magnitude of thermal stresses during the crystal growth exceeds the yield stress, (ii) the degree of degradation in the device performance due to defects depends on their distribution as well as their density. This is because the localized regions of high dislocation density can produce a significantly higher effect on the cell performance compared to a case if the dislocations were uniformly distributed over the cell area, (iii) a knowledge of the defect distribution can allow a better design of the cell and the cell fabrication processes, (iv) determination of the variations in the defect distributions from wafer-to-wafer, and from ingot-to-ingot are essential for material quality control.

The two methods commonly used for determining the dislocation density in semiconductor materials are: X-ray topography and chemical etching. X-ray topography is typically only used for single crystal wafers of low dislocation density. The most commonly used method consists of etching the material in a chemical solution that produces etch pits at the dislocation sites. Subsequently, the etch pits are counted under an optical microscope. This procedure can be extremely tedious and time-consuming for large-area wafers (even with the help of image analysis attachments for the optical microscope). A similar procedure can be applied to polycrystalline silicon substrates. However, care must be exercised in selecting the suitable etches, because many of the etches do not work well for polycrystalline substrates. For example Wright and Dash etches produce etch pits of different shapes and sizes for different orientations (1, 2). A unique etch, known as "Sopori etch," was formulated for polycrystalline silicon to produce etch pits of the same size on all orientations (3). This isotropic etch consists of

HF:CH₃COOH:HNO₃ in the ratio of 36:15:2. The shape of an etch pit produced by this etch depends only on the direction of the dislocation; circular etch pits are formed when the dislocation emerges normal to the surface, and elliptical when the dislocation is at an angle to the surface.

In order to facilitate the defect counting process we have developed a rapid, optical scanning technique that uses scattering from a defect-etched wafer to statistically count dislocations at the surface. This system called Scanning Defect Mapping System (SDMS) can simultaneously produce maps of the grain boundary distributions. This paper describes the principle of this new technique and some applications for semiconductor characterization.

PRINCIPLE OF THE TECHNIQUE

SDMS uses the characteristics of the light from a defect etched sample to recognize and statistically count the dislocation etch pits. When a silicon sample, defect-etched in Sopori etch, is illuminated with a collimated light beam, the reflected light from the etch pits is scattered into a narrow cone of about 20 degrees. Figure 1a shows a reflection pattern due to dislocation etch pits of circular geometry. The shape of the pattern is indicative of the shape of the etch pits. A photograph of the etch pits is shown in Figure 1b. If the etch pits are elliptical, the reflected pattern also shows elliptical geometry.

It has been shown previously that the total integrated light, scattered from an illuminated region of a defect-etched surface, is proportional to the number of dislocation etch pits in that area (4,5). This principle was previously applied to map dislocation density in large-area, single crystal semiconductor wafers. However, the usefulness of this technique for polycrystalline substrates was severely limited. It was found that grain boundary (GB) "grooving," which occurs during defect delineation by chemical etching, produced very strong signals that erroneously indicated a high dislocation density. This problem prevented application of the proposed technique to polycrystalline samples -- a domain where it is expected to be the most valuable. We have recently solved this problem and successfully applied this technique to produce maps of the dislocation distributions in commercial silicon substrates, typically 10 cm x 10 cm in size. Furthermore, this technique is extended to image the grain distribution of the wafer.

In the new approach the reflected light signal, due to the GB scattering, is separated from the dislocation signal. This is made possible by the fact that the divergence of the GB and the dislocation signals is different. The GB scattering occurs with a very large scattering angle and is primarily scattered as line pattern perpendicular to the direction of the GB. Whereas the scattering from the dislocations occurs within a narrower cone as discussed previously (4). These scattering features allow us to retrieve the dislocation and GB information from the scattered and near-specular components of the reflected beam, respectively. In the setup described in the next section, the input and output apertures are adjusted such that the first lobes of the light scattered by the GB's, accompanying the specular reflection, are allowed to emerge from the integrating sphere along with the specular beam. However, the light scattered by the dislocations is captured by the integrating sphere.

SYSTEM CONFIGURATION OF SDMS

Figure 2 is a schematic showing the major components of the SDMS. A light beam from a HeNe laser ($\lambda = 6328 \text{ \AA}$) illuminates the defect-etched sample at normal incidence. The light scattered by the sample is collected by two detectors referred to as the Grain Boundary Detector (GBD) and the Dislocation Detector (DD), shown in Figure 2. These detectors are positioned so as to differentiate between etch pits and grain boundaries as follows. When the laser beam is incident

on a group of etch pits, the light is primarily scattered into a well-defined cone with an angular spread of about 20° . This light is collected by the integrating sphere and is measured by the photodetector DD. The signal from the detector, which measures the integrated light intensity, is proportional to the local dislocation density of the sample. When the laser beam hits a grain boundary, the light is scattered as a one-dimensional streak, elongated perpendicular to the length of the GB. The near-specular component of the scattered light passes through the integrating sphere and is reflected by the beam splitter towards the GBD. The central stop of the annular aperture blocks the axial component of the beam, allowing the off-axis component to focus on the GBD. When the laser beam is incident on a defect-free area of the sample, the light is specularly reflected back out of the integrating sphere, giving no signal on the DD. This beam is then blocked by the annular aperture, thus producing no signal on the GBD either.

The sample is scanned under the light beam using an X-Y stage. The detector signal and the position signals from the X-Y stage are processed by the computer and displayed as defect maps. The SDMS acquires data simultaneously in two modes - the "dislocation" mode, and the "grain-boundary" mode. Electronic circuits are provided to minimize the cross-talk between GB and the dislocation signals. The signals from the two detectors are digitized and fed into a computer along with the x-y position signals from the scanning stage. The data are stored in high speed buffer memory. The electronic hardware and the computer software required for SDMS is developed at NREL. Having the information stored in the computer allows the user to perform detailed analyses of the distributions. Alternatively, the analog signals from the detectors are processed to directly display the defect and GB distributions on a storage oscilloscope.

RESULTS

Figure 3a is a dislocation density map of a 2.5 cm x 2.5 cm commercial solar cell substrate generated by the SDMS. The dislocation densities (corresponding to various colors in the original map) are identified by a gray scale; a calibration sample is used to establish this relationship. Figure 3b shows a grain boundary map of the same sample. Figure 4 is a photograph of the defect-etched sample whose dislocation and GB maps are shown in Figure 3a. In comparing these figures one can clearly note the absence of grain boundaries in Figure 3a and absence of dislocations in Figure 3b.

Due to short times required for mapping dislocations (typically 1 hour for a 10-cm x 10-cm wafer), SDMS can be used routinely to evaluate material quality of commercial photovoltaic silicon substrates. Figure 5a is dislocation map of a polycrystalline silicon wafer typically used for commercial solar cells. A photograph of the defect-etched sample is given in Figure 5b for comparison. It is seen that a large fraction of the total substrate area has very low dislocation density. However, a strong variation in the dislocation distribution seen in this figure is expected to degrade the cell performance due to the fact that heavily dislocated regions can act as "sinks" and shunts for the rest of the device.

In evaluating the material quality by this technique we consider two parameters: the average dislocation density, and the degree of spatial variation in the dislocation density. At this time it is not known how each of these parameters influence the cell performance. These issues are being addressed in our current research work that is being done in collaboration with a number of photovoltaic manufacturing companies.

The dislocation distributions determined by SDMS can also be used on a qualitative basis to determine the locations of high thermal stresses (above the yield stress level) during crystal growth. Figure 6 is a dislocation map of a 10-cm x 10-cm wafer that has a high dislocation density near the wafer edges, indicating a high level of stress at the sides of the casting. It is important to note that this information can only be obtained by extensive defect analyses, such as

by SDMS. Further details of the relationship between defect distributions by SDMS and the crystal growth conditions are given in a forthcoming paper (4).

CONCLUSION

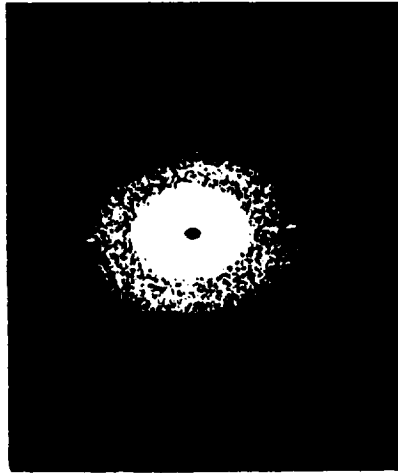
An optical scanning technique for defect mapping of single and polycrystalline silicon substrates has been developed. This technique uses optical recognition of dislocations and grain boundaries and performs statistical counting of dislocation etch pits. These features make this technique rapid and accurate compared to any other technique currently available. Furthermore, the cost of the instrument is very low compared to other instruments based on imaging and electronic recognition. Although we have demonstrated use of this technique for silicon, this is also applicable to other semiconductors.

ACKNOWLEDGMENT

The author would like to acknowledge the contributions to this work by Robert Murphy, Craig Marshall, and Larry Allen of NREL. This work was supported by NREL Technology Maturation Fund and by the United States Department of Energy under DOE Contract # DE-AC02-83CH10093.

REFERENCES

- [1] W. C. Dash, J. Appl. Phys. **27**,1193(1956).
- [2]. M. W. Jenkins, J. Electrochem. Soc., **124**, 734(1976).
- [3] B. L. Sopori, J. Electrochem. Soc. , **131**, 667(1984).
- [4] B. L. Sopori, Appl. Optics, **22**, 4676(1988).
- [5] B. L. Sopori, J. Electrochem. Soc., **135**, 2601(1988).



(a)



(b)

Figure 1a. Photograph of the reflection pattern from circular etch pits

Figure 1b. Photograph of the circular etch pits

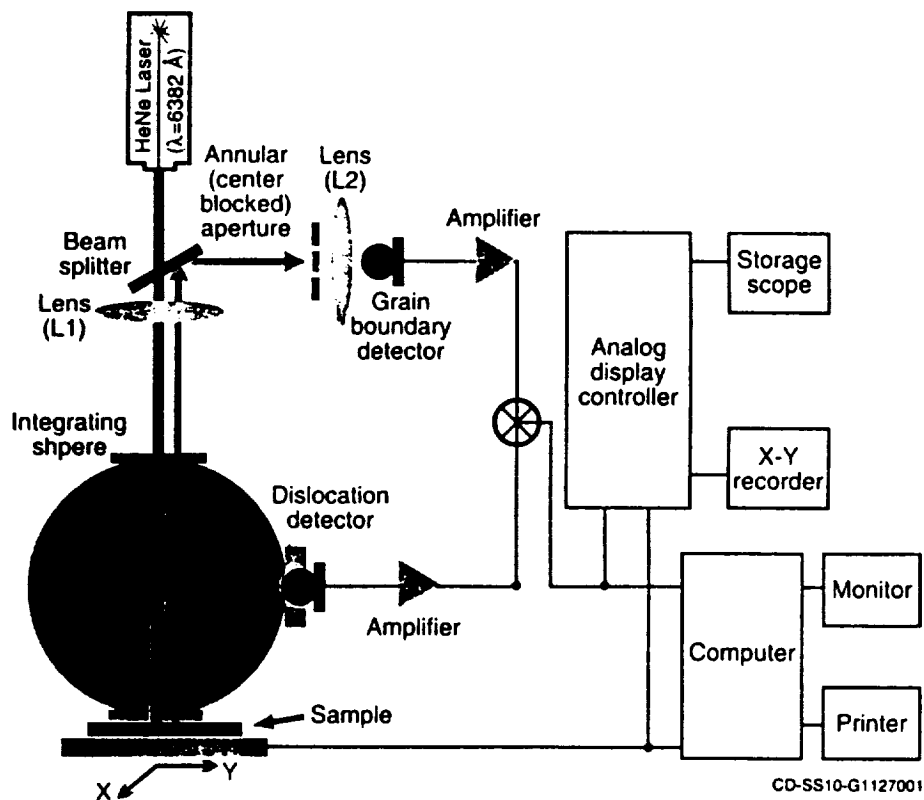


Fig. 2. A schematic of the Scanning Defect Mapping System

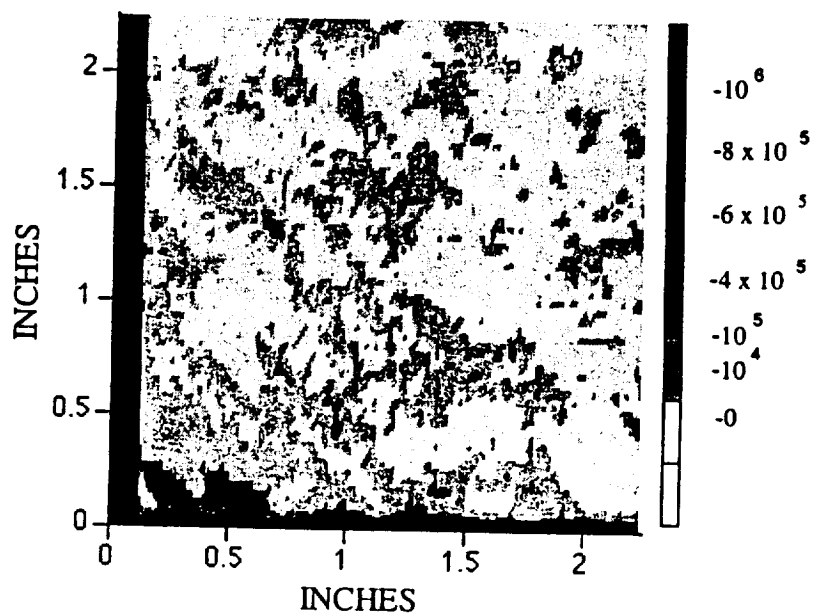


Figure 3a. Dislocation map of a 5-cm x 5-cm polycrystalline silicon substrate (shown in Fig. 4) produced by SDMS



Figure 3b. Grain-boundary map of the 5-cm x 5-cm sample shown in Fig. 4



Figure 4. Photograph of the defect-etched sample used in Fig. 3

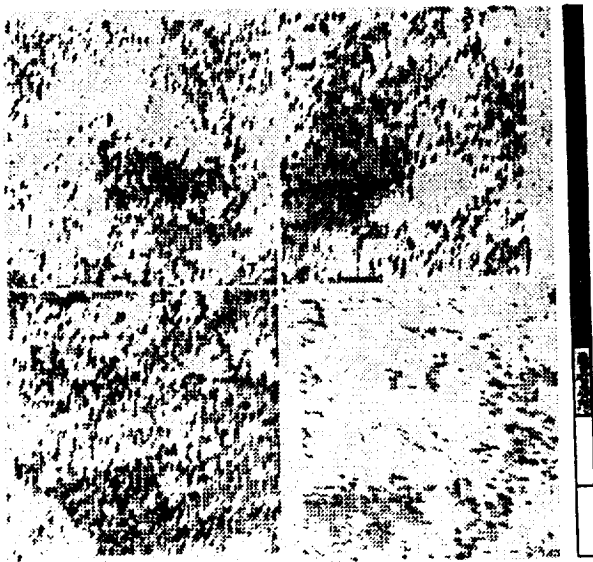


Figure 5a. Dislocation map of a 10-cm x 10-cm commercial silicon wafer showing non-uniform dislocation distribution

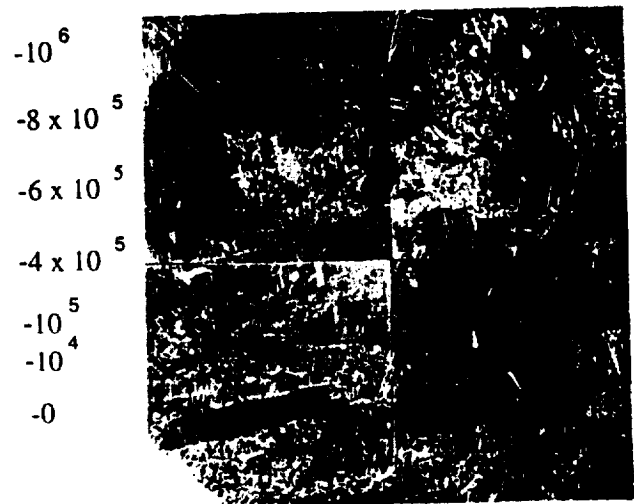


Figure 5b. Photograph of the defect-etched sample corresponding to the dislocation map of Fig. 5a

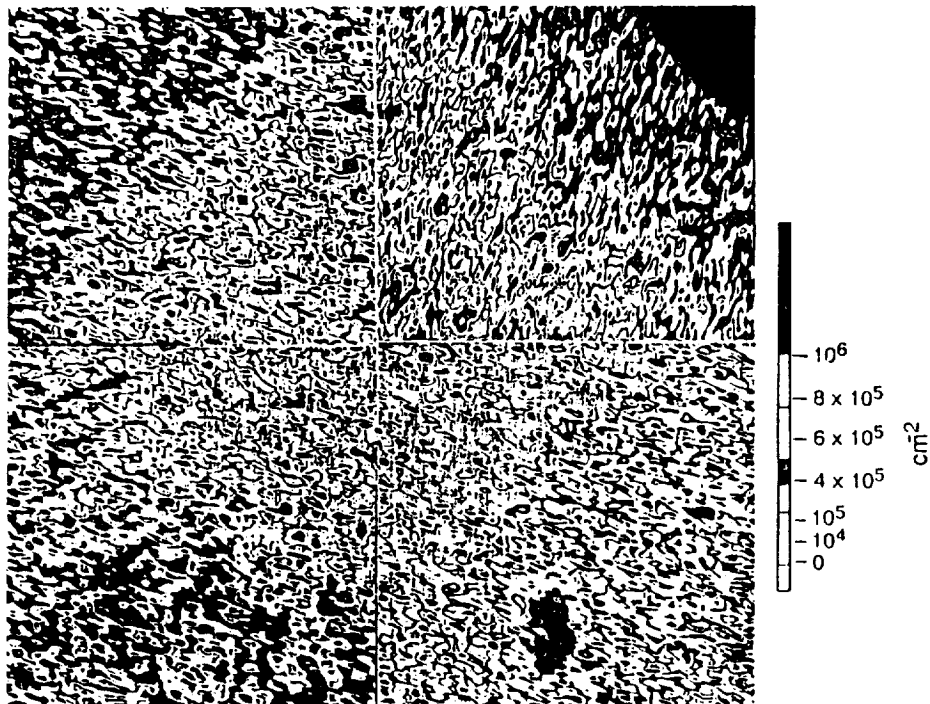


Figure 6. Dislocation density map of a 10-cm x 10-cm corner section of a cast wafer indicative of a high thermal stress at the edges of the casting as manifested by a high dislocation density around the edges.

NEUTRAL ION SOURCES IN PRECISION MANUFACTURING

Steven C. Fawcett

Optical Systems Branch, EB53, National Aeronautics and Space Administration
Marshall Space Flight Center, AL 35812

Thomas W. Drueding

Aerospace and Mechanical Engineering, Boston University
Boston, MA 02215

ABSTRACT

Ion figuring of optical components is a relatively new technology that can alleviate some of the problems associated with traditional contact polishing. Because the technique is non contacting, edge distortions and rib structure print through do not occur. This initial investigation was aimed at determining the effect of ion figuring on surface roughness of previously polished or ductile ground ceramic optical samples. This is the first step in research directed toward the combination of a pre-finishing process (ductile grinding or polishing) with ion figuring to produce finished ceramic mirrors. The second phase of the project is focusing on the development of mathematical algorithms that will deconvolve the ion beam profile from the surface figure errors so that these errors can be successfully removed from the optical components. In the initial phase of the project, multiple, chemical vapor deposited silicon carbide (CVD SiC) samples were polished or ductile ground to specular or near-specular roughness. These samples were then characterized to determine topographic surface information. The surface evaluation consisted of stylus profilometry, interferometry, and optical and scanning electron microscopy. The surfaces were ion machined to depths from 0-5 μm . The finished surfaces were characterized to evaluate the effects of the ion machining process with respect to the previous processing methods and the pre-existing subsurface damage. The development of the control algorithms for figuring optical components has been completed. These algorithms have been validated with simulations and future experiments have been planned to verify the methods. This paper will present the results of the initial surface finish experiments and the control algorithms simulations.

INTRODUCTION

In the new precision fabrication facility currently being implemented at Marshall Space Flight Center (MSFC), the ion figuring technology is being evaluated to complement traditional high precision material removal techniques. The use of neutral ion beams to remove selected material has been developed and is currently in use at a few commercial installations and research facilities [1-10]. These implementations typically use the process for final figure correction on meter class optical components. The new facility at MSFC will develop and evaluate this technique for producing the optics required on many of NASA's missions. The initial work focuses on the production of centimeter scale optics to achieve the extremely tight tolerances in a reasonable time. After the development work has been completed, the fabrication procedures will be optimized for the particular requirements of various missions. This technology is extremely important to advance the fabrication techniques used in the production of high precision components. The process has a wide variety of commercial applications in the areas of optics manufacturing as well as other potential uses in fields where components are required with precision figure and surface finish tolerances. The use of neutral ion sources for material fabrication is probably one of the key technologies for manufacturing in the next century.

Loose abrasive polishing and ductile regime grinding are two fabrication processes that can provide optical quality surface finish on a glass or ceramic mirror. Polishing usually involves loose abrasive particles working between a flexible pad and the mirror material, creating the desired surface characteristics by either frictional abrasion or chemical reactions or both. A newer fabrication procedure that can potentially be used for optical surface fabrication is ductile regime grinding [11-14]. In this technique, the substrate is formed on a precisely controlled fixed abrasive grinding machine. The process parameters are controlled to allow for material removal by plastic deformation instead of brittle fracture. Both polishing and ductile grinding can produce the angstrom level surfaces required by precision optical components. It should be noted that the ion figuring technique require that the mirror segments have optical quality surface finishes prior to figuring since the ion machining can not currently be utilized to improve the finish of components. Polishing and ductile-regime grinding will be evaluated as preliminary fabrication technologies, in which the contour accuracy of the component is achieved to within 1-2 μm of the final specification, and the component is finished to the final surface roughness specification. Subsequent ion-figuring will then be used to improve the contour accuracy while maintaining angstrom-level roughness.

The use of neutral ion beams to remove selected material has been developed for final figure correction on meter class optical components [5-8]. The technique has been developed for this size application with commercially available, three to five centimeter ion sources. These sources are utilized for the correction of large and middle spatial frequency errors on optical components. The procedure uses an interferometric contour subtracted from an ideal contour to produce a map of the figure errors. The process itself is based on a momentum transfer from previously ionized molecules of inert gas. This results in molecular level sputtering of the substrate material. The basic operation is depicted in Figure 1.

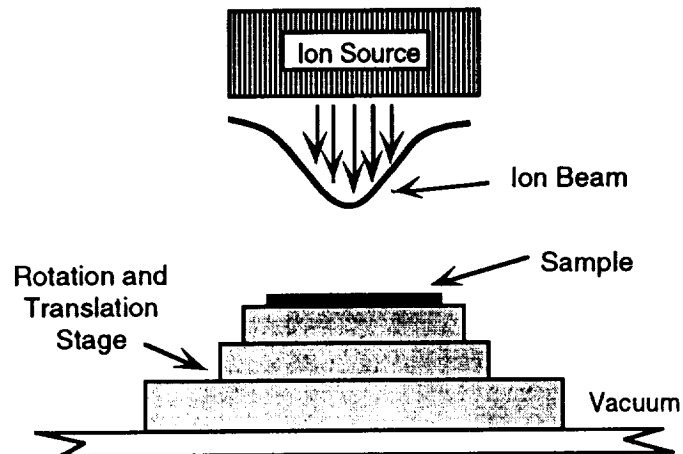


Figure 1 Schematic of ion figuring process. A semi-Gaussian ion function imparts the final figure contour on the substrate by computer controlled translation and rotation.

The new research facility at MSFC will focus on using the technique for the figuring of centimeter scale optics [15]. This facility was constructed around a surplus sputtering chamber obtained for this program. This chamber was retrofitted with a 3 cm, Kaufman filament type ion source [16]. This source is driven by a programmable power supply which can provide current densities to 10 mA/cm^2 . The basic system is shown in the computer model of Figure 2.

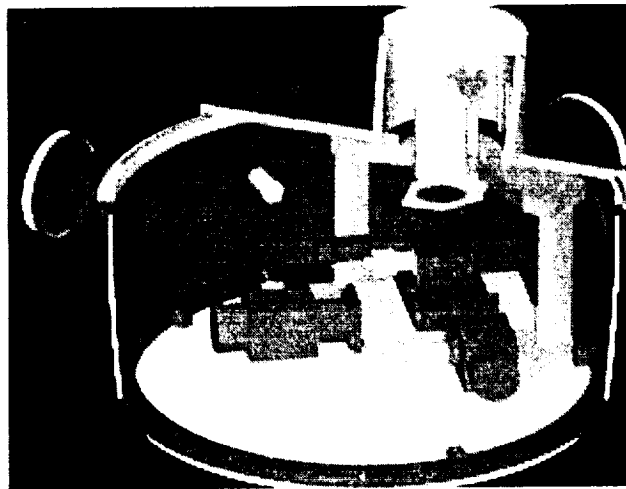


Figure 2 Solid model of the ion figuring system components.

The initial experiments were designed to evaluate the machining characteristics of Chemical Vapor Deposited silicon carbide (CVD SiC). The polished samples were 3 cm diameter and the ductile ground samples were 6 mm wide by 25 mm long rectangular pieces. The polished samples were processed by lapping with an oil based diamond compound. The ductile ground samples were obtained from the same batch as the polished samples. They were mounted on steel sample holders and ground in the ductile regime with a fixed abrasive diamond grinding wheel.

An experiment was performed to determine the beam current profile and hence the expected material removal rate on the CVD SiC samples. In this experiment, a polished, circular sample was located beneath the ion beam so that the edge of the sample would correspond with the point of maximum beam intensity. One half of the sample was masked to provide a reference surface for step height measurements. This process is depicted in Figure 3. Throughout these experiments, the ions were produced from Argon gas. After the sample was ion machined for a specified period, the relative height differences between the surfaces was measured with a Talysurf stylus profilometer and normalized with respect to time. Figure 4 shows a plot of the measured machining rate versus the distance from the beam center.

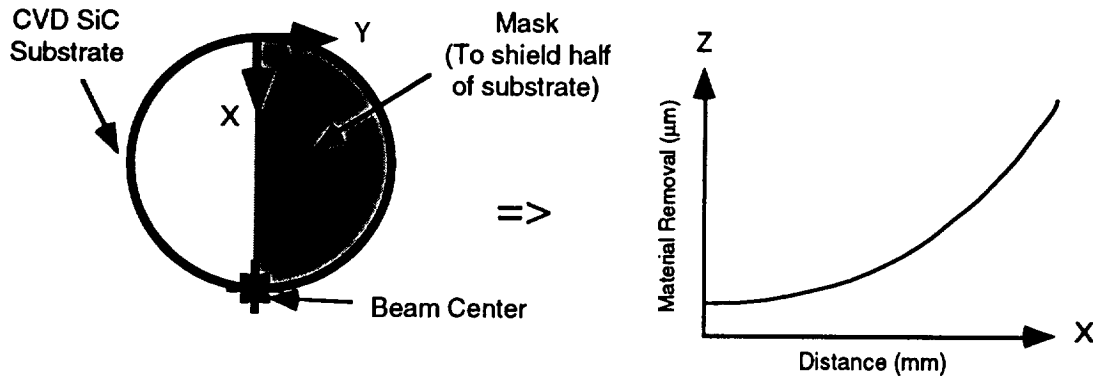


Figure 3 Masking technique for determining machining rate.

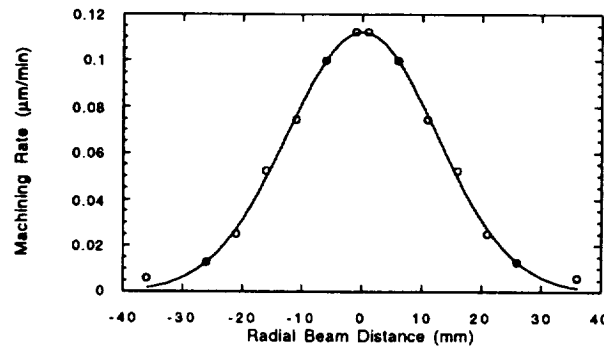


Figure 4 Machining profile of 3 cm ion source.

Once the samples were polished or ground, surface characterization measurements were made. Unmagnified, visual examination revealed no scratches or pits in any of the samples. Optical microscopy to 400 X only revealed an occasional fracture location and grooves on the ground samples caused by the grinding crossfeed, especially on sample D1. The samples were viewed with a scanning electron microscope (SEM). There were small ($\approx 100 - 200$ nm diameter) discontinuities visible in the SEM micrographs of the ground samples. On all the polished samples, small (~ 100 nm wide) scratches became visible. All the samples were measured to determine the surface roughness parameters prior to ion machining. It was determined that the surfaces roughness of the samples was below the noise floor of the Talysurf profilometer. The initial measurements were instead made on a WYKO optical profilometer to determine the surface roughness. Ductile ground sample "D4" was also measured by Mark Gerchman of Rank Taylor Hobson, Inc. on a Talystep stylus profilometer for verification. Multiple measurements were made for each sample and the results averaged. The RMS values from the WYKO data for an area approximately $235 \times 235 \mu\text{m}$ square ranged from $7.56 \pm 2.07 \text{ \AA}$ to $12.77 \pm 4.65 \text{ \AA}$ for the polished samples and from $12.87 \pm 4.94 \text{ \AA}$ to $52.25 \pm 34.51 \text{ \AA}$ for the ductile ground samples. The Talystep data for sample "D4" gave an average roughness of $13.14 \pm 5.24 \text{ \AA}$ RMS for a $150 \mu\text{m}$ linear trace.

SURFACE FINISH RESULTS

After the initial surface finish evaluation, the samples were ion machined in a fixed position with respect to the beam, so that the total depth profile at various positions on the sample surface could be correlated with the ion

machining depth profile plotted in Figure 4. These samples were then examined to determine the effect of the ion sputtering process. During machining, the samples could be viewed through a transparent port in the vacuum chamber. This examination revealed an interesting phenomenon: previously hidden damage on the polished samples became highly visible after only a few minutes under the beam. After machining, some of the samples showed extensive damage when examined under an optical microscope. The extent of this damage increased with an increase in ion machining depth. Optical micrographs revealed hemispherical pits uncovered by the ion machining process on the polished samples. The pits are believed to have been formed at the sites of sub-surface damage caused by the loose abrasive polishing with the larger grit diamond. The size of the pits are approximately 10 to 20 μm diameter which corresponds to the fracture size expected from the initial polishing stage with 15 μm diamond. Because of the lack of a controlled polish, it is probable that the fractures from previous grit sizes was not completely removed in subsequent steps and were left as sub-surface damage. These damage sites were then uncovered by the ion figuring. The micrographs of Figure 5 appears to verify the hypothesis that the pits are formed in the areas of sub surface fracture. Figure 5 shows a line of these hemispherical pits formed on an uncovered sub-surface scratch. These pits apparently follow a scratch line where fractures occur under the stresses induced while polishing with the larger grit sizes.

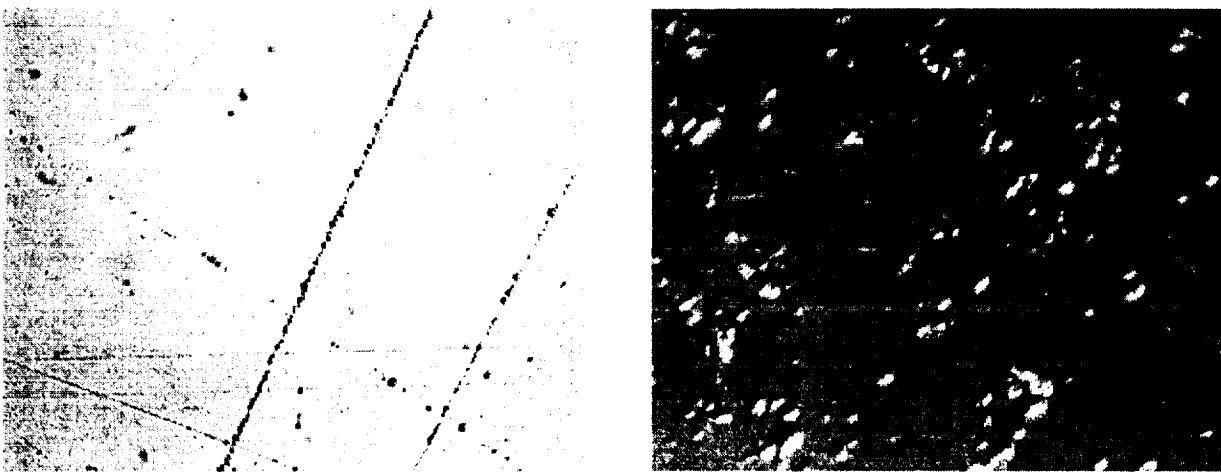


Figure 5 Optical micrographs of subsurface fracture revealed by ion machining on a polished CVD SiC sample (200 X left and 400 X right).

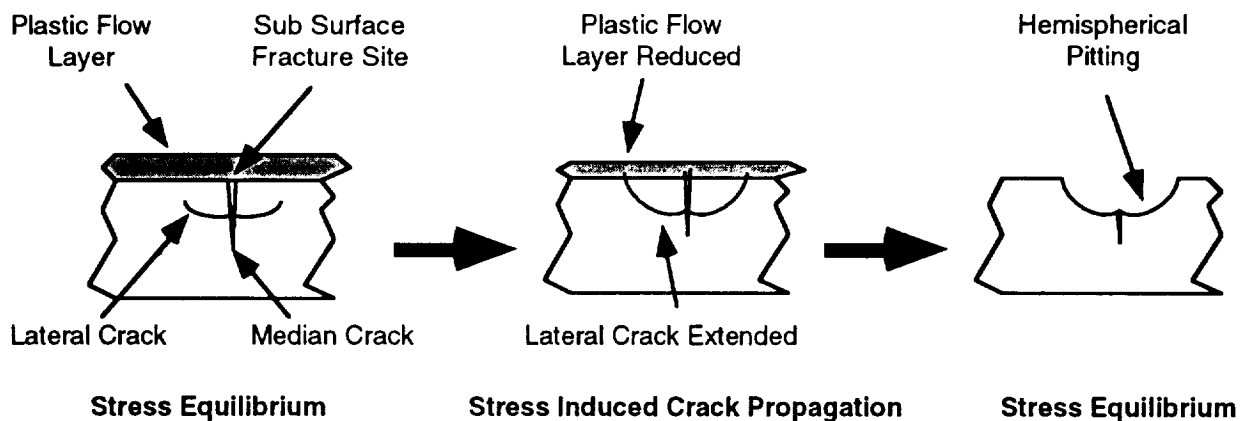


Figure 6 Schematic of the proposed crack propagation process. The layer that provided stress equilibrium is removed and the sub-surface crack propagates eventually reaching the surface.

It is believed that the pitting sites are formed when the lateral cracks caused by the forces of the individual diamond grits are uncovered. The lateral cracks may potentially propagate due to the removal of the plastic flow layer which initially provided the stresses required to maintain equilibrium. Figure 6 demonstrates this idea schematically. When the plastic layer is removed by the ion sputtering process, the equilibrium condition is modified by removal of the

compressive surface layer with the resulting increase in tensile stress below the surface, and the cracks grow. When the surface layer meets the extended lateral crack, the material above the crack is expelled, (releasing the residual stress) to reveal the hemispherical pits. The thermal loading inherent with the ion figuring process may be a contributing factor. The areal density of these fracture sites appears to directly correlate with the density of the ion beam current and hence with the machining depth.

The ground samples were also evaluated optically. The samples ground in the ductile regime showed only a few fracture sites in localized areas of the samples. These areas appeared to correspond with sites which may have undergone some scratching during the grind. However, most of the areas were free of fracture. The pits shown range from 3 to 8 μm diameter which is similar to the diamond grit size of 4 to 8 μm . The brittle ground sample "D1" showed some hemispherical pitting similar to the polished samples. Surface roughness measurements were also made on the figured samples. Prior to machining, six measurements were made at random locations for each of the samples with the WYKO optical profilometer. These results were averaged and are shown in Table 1.

Sample	WYKO Data (\AA)
P 1	7.56 \pm 2.07
P 2	7.95 \pm 0.62
P 3	10.5 \pm 5.31
P 4	12.77 \pm 4.65
D1	52.25 \pm 34.51
D2	15.78 \pm 4.02
D3	12.87 \pm 4.94
D4	17.45 \pm 2.07

Table 1 RMS surface finish of CVD SiC samples (WYKO measurement area approximately 235 x 235 μm).

After ion machining, the samples were again measured with the optical profilometer. Those samples exhibiting significant pitting (including all of the polished samples and two of the four ground samples) could not be measured with the WYKO optical profilometer. The steep slope of the pit edges exceeded the measurement capacity of the optical profilometer. Data was obtained from the WYKO for three of the ground samples that had relatively few pits (ductile ground samples D2, D3 and D4). The RMS surface finish measurements made over the entire surface were averaged to obtain 15.77 \pm 2.14 \AA for sample D2, 17.98 \pm 4.06 \AA for sample D3 and 23.01 \pm 6.95 \AA for sample D4. The area of measurement for the samples was approximately 470 x 470 μm for D2 and 235 x 235 μm for D3 and D4. This surface roughness revealed no significant increases as a function of machining depth. Those samples that could not be measured with the optical profilometer were instead measured using a Talysurf stylus profilometer. This data at each ion machined depth was collected in four linear, 1 mm traces of the samples with no filtering. The RMS roughness average of each set of four traces is plotted against ion machining depth in Figures 7 and 8. Figure 7 shows the surface roughness as a function of ion machining depth for the polished samples. There is an obvious increase in the surface roughness at the greater machining depths. This roughness increase is primarily due to the pitting that occurred in these polished samples.

Sample "P4" does not show the same increase in roughness as the other polished samples. This sample was polished for a longer time than the others, suggesting that the observed pitting may be the result of rough-machining steps prior to polishing, and that the polished samples were not polished long enough to remove all prior damage. Figure 8 shows the surface roughness measurements for the ground sample D1. The other ground samples exhibited roughness that was below the noise floor of the Talysurf profilometer. The sample, D1, which was ground in the brittle regime, exhibits the same trend of increasing roughness with ion machining depth that was observed in the insufficiently polished samples. Before ion machining, D1 had a roughness of 52 \AA RMS. Also this sample had an abundance of damage sites that became visible under the optical microscope after ion machining. It should be noted that the vertical resolution of the Talysurf is 10 nm and results in the absence of measurements below this threshold in Figures 7 and 8. The pitting exhibited on the polished samples and the ground sample D1 prevented Wyko optical profilometry measurements for these samples.

DISCUSSION OF SURFACE FINISH EXPERIMENTS

From these results, it seems apparent that ion machining acts to expose, and possibly amplify, any pre-existing subsurface damage in specularly finished CVD SiC. This exposure of damage causes significant roughening of the surface, diminishing its optical quality. Both pre-finishing techniques, grinding and polishing, can be used to

generate specular surfaces. However, the *roughness* of the finished surface is not indicative of the extent of subsurface damage. All of the polished samples, and three of the four ground samples exhibited RMS roughness before ion machining that were below 20 Å. After ion machining to a depth of 5 μm , those samples having significant and visually apparent subsurface damage exhibited increases in RMS roughness to as much as 1400 Å, while those samples that showed no evidence of subsurface damage exhibited almost no increase in RMS roughness.

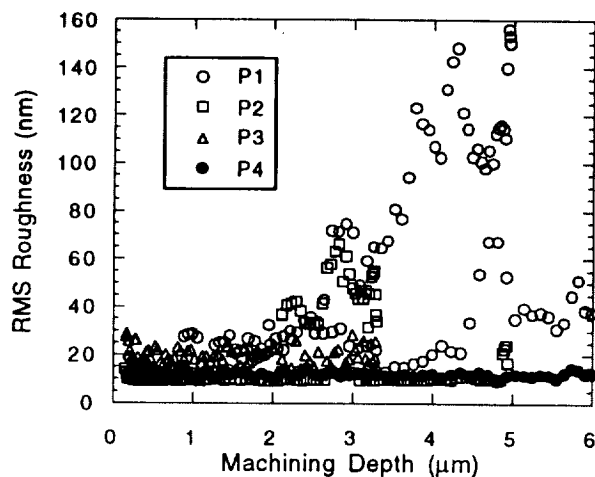


Figure 7 Surface roughness variation with ion machining depth for polished samples.

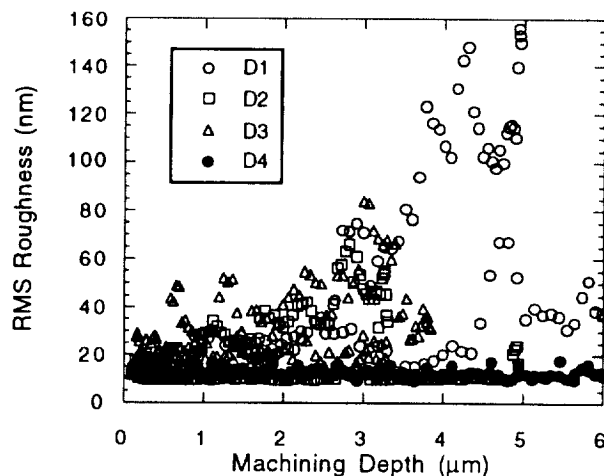


Figure 8 Surface roughness variation with ion machining depth for brittle ground sample D1.

The subsurface damage uncovered by ion machining might have been the result of rough machining processes prior to polishing or grinding, or they might have been introduced by these finishing processes themselves. The evolution of roughness in the polished samples, however, provides evidence suggesting that prior machining processes are responsible: the polished sample exhibiting the least amount of subsurface damage had been polished for at least twice as long as the other samples but no verifiable material removal measurements were ever made during the polishing. If the polishing process itself were a primary contributor to the subsurface damage, this sample would have been expected to exhibit at least as much damage as the others. It is perhaps not surprising that the subsurface damage from previous rough machining was not completely removed in the polished samples: the large value of hardness exhibited by SiC causes it to have an exceptionally low polishing rate. These polishing

experiments were conducted with diamond abrasive on a Kemet lap and with B₄C abrasive on a cast iron lap. Although it is possible to increase the polishing rate by polishing at a higher pad pressure, it is also possible that the increased polishing grain loading would itself result in polish-induced subsurface damage.

Ductile-regime grinding appears to be a good way to obtain the two prerequisites for ion machining: a specular finish and the absence of subsurface damage. It has been demonstrated in these experiments that it is possible to grind SiC to RMS roughness in the 10-20 Å range, without introducing any subsurface damage. If the grinding process is not carried out in the ductile regime, however, then the RMS roughness of the sample will increase with subsequent ion machining. An interesting result of these experiments is the observation that ion machining acts to expose subsurface damage in SiC. By itself, this process may be useful as a locally-destructive diagnostic technique for evaluating the subsurface damage in a specular component, in the same way that HF acid etching is used to evaluate subsurface damage in glass. Since acid does not readily etch SiC, ion machining can serve as a clean, efficient alternative for post-processing measurement of subsurface damage in ceramic components. These experiments provide strong evidence that ceramic mirrors could be produced by pre-finishing in a ductile machining process (either ductile-regime grinding or polishing), followed by ion machining. A distinct advantage of such a fabrication technique is that, at the end of the ion milling process, the mirror surface is atomically clean, and it is in a vacuum chamber. This means that the surface is ready for any subsequent coating processes that are to be performed on the mirror, processes that generally require atomic cleanliness and a vacuum environment anyway.

FIGURE CONTROL ALGORITHMS

An important step in the fabrication of an optical component involves the expensive and time consuming process of imparting a precise contour on the optic. Ion beam figuring provides a highly deterministic method for the final precision figuring of optical components with advantages over conventional methods. The repeatability of the process allows the possibility of single step figuring, resulting in significant time and cost savings. Unlike grinding, polishing and lapping, ion figuring is non-contacting and so avoids several problems including: edge effects, tool wear, and loading of the work piece. The work discussed here is directed toward the development of control algorithms for ion machining of small (≤ 10 cm diameter) optical components. The development of the algorithms has proceeded in three steps: 1) the development of control algorithms; 2) the use of a simulation to verify the methods and to identify important issues that are critical for the proper operation of the system 3) the integration of the controlling algorithms with the other components. Experimental testing of the system will then follow and includes the figuring of chemically vapor deposited silicon carbide (CVD SiC) and fused silica samples. An emphasis will be placed on CVD SiC that have been previously fabricated by ductile-regime grinding.

The ion figuring process involves bombarding a surface with a temporally and spatially invariant beam of neutralized particles, typically argon. Material on the surface is consequently sputtered away. The rate and distribution (profile) of material removal by the ion beam is kept constant (as shown in Figure 4) and specific figures or corrections are achieved by rastering the fixed-current beam across the workpiece at varying velocities. The figuring process relies on obtaining an accurate figure error map of the workpiece surface. The solution process consists of choosing a pattern the beam traverses and determining the rastering velocities along that path. The removed material can be computed by convolving the ion beam's fixed removal profile and the rastering velocity as a function of position. The determination of the appropriate rastering velocities from the desired removal map and the known ion beam removal profile is therefore a *deconvolution* process. A unique method of performing this deconvolution was developed for the project based on research on singular expansions of filters [17]. The solution to a convolution equation can be expressed in terms of derivatives of the known function and inverse moments of the filter. The deconvolution algorithm requires information on the moments of the ion beam removal profile and a series expansion of the surface map of the work piece. A unique and stable solution is then directly available. The assumption in this algorithm is that the expansion represent the desired removal map. There is no special consideration to the geometry of the workpiece in the x-y plane. The solution provided is exact so the accuracy of expansion fit to the desired removal is the measure of the solution's accuracy.

SIMULATION RESULTS

A simulation of the process was developed to verify the algorithms and to identify important issues and parameters that are critical for the proper operation of the system. All simulations involve data from an actual 8 cm flat sample. The process of figuring the sample to ideal flat and spherical contours is then attempted by performing a discrete convolution of the ion beam removal profile and the computed solution. The resulting reduction in the RMS difference between the actual surface contour and the desired surface contour is used as the measure of the efficiency of

the process. The sample has an initial error of 234 nm RMS from a flat plane, and 113 nm RMS from a 400 m radius sphere. The expansions representing the error contours (removal map) are polynomial series. The higher the order of the series, the more accurate the solution. Initial experiments were performed to evaluate the effect of the order on the performance of the system. The number of terms used in the series expansion representing the removal map was increased and corresponding increases in the efficiency of the process were observed. The results are shown in Figure 11 as the final RMS deviation from an idealized flat plane.

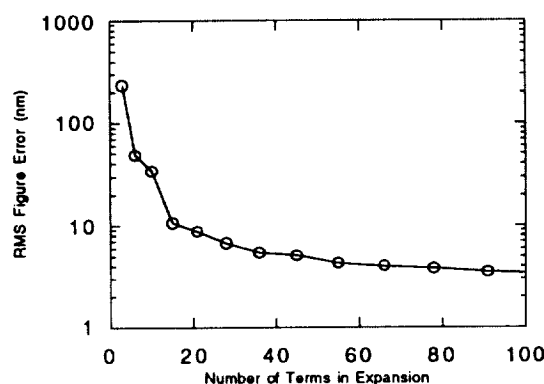


Figure 11 Results of simulation to determine effects of expansion order.

In Figure 12, the first point represents the initial RMS difference at the three expansion terms required to define a plane. The final RMS difference is directly related to the accuracy of the expansion fit. Using fifteen terms the error was reduced to 10 nm RMS and after one hundred terms to 3 nm RMS. Above approximately fifteen terms there is no significant increase in figure accuracy and 10 nm RMS is well below the accuracy of the original data. Other simulations were performed to determine the effects of systematic and random errors. The errors were introduced through the variations in the model of the beam function. The ideal beam function is modeled as a Gaussian distribution defined by two parameters: the width (λ) and the amplitude (Γ).

$$H(x,y) = \Gamma \exp\left(-\frac{x^2 + y^2}{\lambda^2}\right)$$

The sensitivity of process to systematic and random errors in the parameters is observed in Figures 12, 13 and 14. These plots depict the improvement in the RMS difference from the desired figure. This information provides a quantified measurement of the relative sensitivity to errors in the beam parameters as well as the positioning of the beam with respect to the part. Experiments are planned to verify the simulations and to provide information on the accuracy and stability of the beam parameters and positioning.

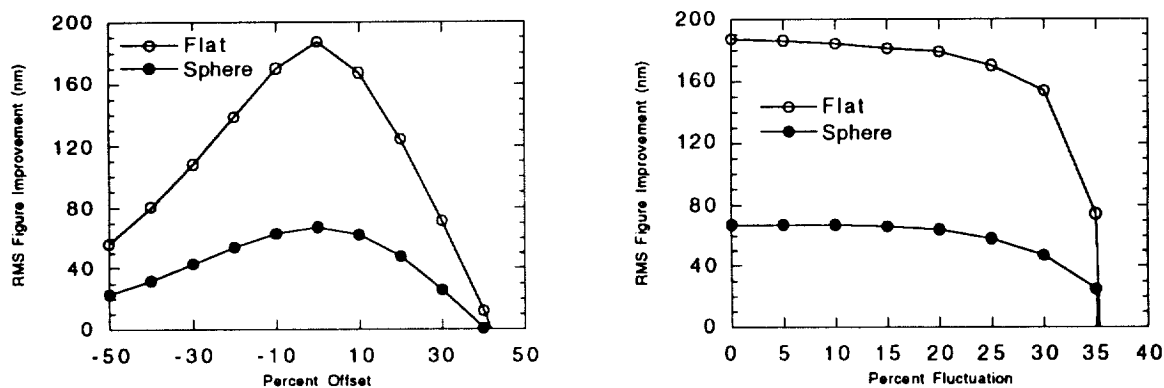


Figure 12 Simulated improvement in RMS figure error shown as a function of systematic and Normally distributed random error as a percentage of the modeled λ .

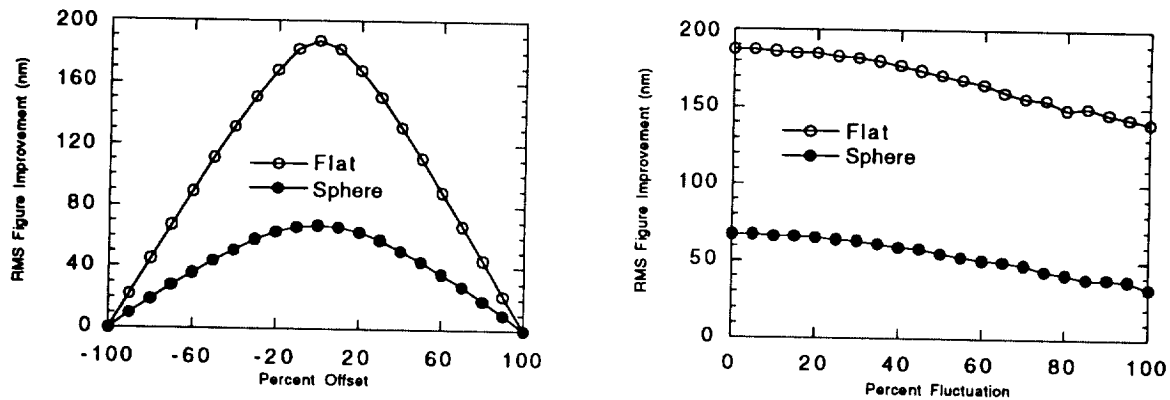


Figure 13 Simulated improvement in RMS figure error shown as a function of systematic and Normally distributed random error as a percentage of the modeled Γ .

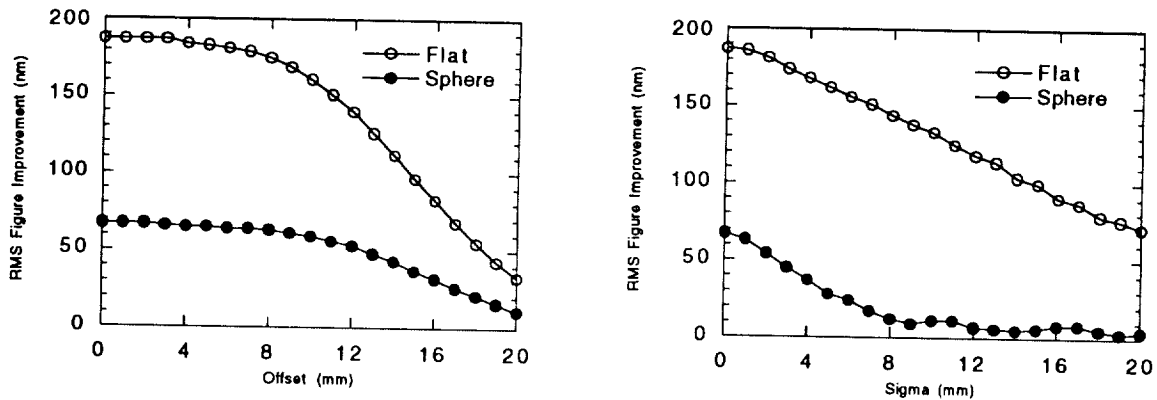


Figure 14 Simulated improvement in RMS figure error shown as a function of systematic and Normally distributed random position error.

CONCLUSIONS

It has been demonstrated that a feasible process for fabricating ceramic mirrors is to employ a deterministic ductile-grinding operation followed by a final ion beam contouring operation. A small ion machining facility has been established, and processing experiments on CVD SiC have been carried out. Removal rates for the ion beam have been characterized. The removal profile is largely gaussian in shape. CVD SiC samples were polished or micro-ground in preparation for final ion machining. Most of these pre-finished samples exhibited roughness in the 10-20 Å range. It was found that the success of the ion machining process was directly influenced by the existence of subsurface damage: samples containing subsurface damage suffered rapid deterioration in surface quality as a result of the ion machining. The observed sub-surface damage in polished samples is likely to be the residual effect of previous machining processes, rather than the direct result of polishing itself. As has been shown in previous work by many researchers, ductile grinding can produce damage-free, optical quality surfaces in SiC. Initial experiments revealed that high quality, ductile ground surfaces do not significantly roughen after removal of up to 5 μm of material by ion machining. The deconvolution algorithm for figuring optical components has been successfully developed and verified with simulations. Experiments are currently underway to validate these simulations on actual SiC and fused silica optical flats.

The proposed processing technique shows promise as a viable alternative to conventional fabrication techniques in the production of mirror components and other precision surfaces. Moreover, the results of the experiments reported here indicate that ion machining can serve as an indicator of the existence and extent of subsurface damage in

specularly finished ceramic components. Because the ion sputtering rate is dependent on material properties and atomic species, this process can be utilized for producing surface features in many different substrates. This technique is successfully utilized on a commercial basis for texturing surfaces on a relatively microscopic scale. The future of the process will be in producing macroscopic features such as optical surfaces. Another use can be found in sharpening diamond tools by the preferential etching along crystal orientation. Recent experiments performed at MSFC also used the process to produce conical ends on gradient index optical fibers by the differential etching of the varying density material. The future will only bring increasing uses for this technique. Once the groundwork has been laid, commercialization of the process will naturally follow.

ACKNOWLEDGMENTS

This work was sponsored and funded by NASA. The authors would like to express their gratitude to the following people for their support of the project; Robert Rood, Thomas Bifano, James Bilbro, Edward Montgomery, Whit Brantley, Charles Jones, Joseph Randall and Scott Wilson. Also, special thanks to Charles Griffith and Raj Khanijow who performed the tedious job of polishing the samples and to Staffan Persson who designed the support fixtures for the ion source.

REFERENCES

1. A.J. Gale, "Ion Machining of Optical Components", Optical Society of America Annual Meeting Conference Proceedings, November 1978.
2. J.R. McNeil and W.C. Herrmann, Jr., "Ion Beam Applications for Precision Infrared Optics", *Journal Vacuum Science and Technology*, 20 (3), pp. 324-326, March 1982.
3. J.R. McNeil, S.R. Wilson and A.C. Barron, "Ion Beam Figuring for Rapid Optical Fabrication", *Optical Society of America Workshop on Optical Fabrication and Testing*, pp. 62-67, October 1986.
4. S.R. Wilson, D.W. Reicher and J.R. McNeil, "Surface Figuring Using Neutral Ion Beams", *Advances in Fabrication and Metrology for Large Optics*, SPIE Vol. 966, pp. 74-81, August 1988.
5. L.N. Allen, and R.E. Keim, "An Ion Figuring System for Large Optic Fabrication", *SPIE Vol. 1168*, pp.33-50, August 1989.
6. T.J. Wilson, "New Technologies Fabricate Large Aspheres", *Laser Focus World*, Vol. 26(9), pp. 111-123, September 1990.
7. L.N. Allen, R.E. Keim T.S. Lewis and J. Ullom, "Surface Error Correction of a Keck 10m Telescope Primary Mirror Segment by Ion Figuring", *SPIE*, Vol. 1531, July 1991.
8. L.N. Allen, J.J. Hannon and R.W. Wambach "Final Surface Error Correction of an Off-Axis Aspheric Petal by Ion Figuring", *SPIE*, Vol. 1543, July 1991.
9. S.R. Wilson, and J.R. McNeil, "An Update on Ion Beam Figuring", *Optical Society of America Workshop on Optical Fabrication and Testing*, October 1992.
10. C.M. Egert, "Roughness Evolution of Optical Materials Induced by Ion Beam Milling", *SPIE Vol. 1752*, 1992.
11. J. Yoshioka, K. Koizumi, M. Shimizu, H. Yoshikawa, M. Miyashita, and A. Kanai, "Ultraprecision Grinding Technology for Brittle Materials: Application to Surface and Centerless Grinding Processes", *Milton C. Shaw Grinding Symposium, ASME*, Vol. 16, pp. 209-227, 1985.
12. Bifano, T.G. T.A. Dow and R.O. Scattergood, "Ductile Regime Grinding: A Technology for machining Brittle Materials", *Journal of Engineering for Industry*, Vol. 113, pp. 184-189, 1991.
13. T.G. Bifano, and S.C. Fawcett, "Specific Grinding Energy as an In Process Control Variable for Ductile Regime Grinding", *Precision Engineering*, Vol. 13(4), pp. 256-262, 1991.
14. T. G.Bifano, Kahl, W. K., and Yi, Y., "Fixed-Abrasive Grinding CVD Silicon Carbide Mirrors," submitted to *Precision Engineering*, December, 1992.
15. S.C. Fawcett, and R.W. Rood, "Ion Figuring System for Segmented, Adaptive Optics", *American Society for Precision Engineering Annual Meeting Conference Proceedings*, pp. 244-247, October 1992.
16. H.R. Kaufman, P.D. Reader and G.C. Isaacson, "Ion Sources for Ion Machining Applications", *AIAA Journal*, Vol. 15(6), pp. 843-847, June 1977.
17. R. G. Hohlfeld, J. I. F. King, T. W. Drueding, and G. v. H. Sandri, "Solution of Convolution Integral Equations by the Method of Differential Inversion", *SIAM Journal of Applied Mathematics*, 53(1), 1993.

2511
P-13
HIGH POWER DIODE LASERS FOR SOLID-STATE LASER PUMPS

Kurt J. Linden
Director, Commercial Products and Services
Optoelectronics Division
Spire Corporation
One Patriots Park
Bedford, MA 01730-2396

Patrick N. McDonnell
Vice President and General Manager
Optoelectronics Division
Spire Corporation
One Patriots Park
Bedford, MA 01730-2396

ABSTRACT

The development and commercial application of high power diode laser arrays for use as solid-state laser pumps is described. Such solid-state laser pumps are significantly more efficient and reliable than conventional flash-lamp pumps. This paper describes the design and fabrication of diode lasers emitting in the 780 - 900 nm spectral region, and discusses their performance and reliability. Typical measured performance parameters include electrical-to-optical power conversion efficiencies of 50%, narrow-band spectral emission of 2 to 3 nm FWHM, pulsed output power levels of 50 watts/bar with reliability values of over 2 billion shots to date (tests to be terminated after 10 billion shots), and reliable operation to pulse lengths of 1 ms. Pulse lengths up to 5 ms have been demonstrated at derated power levels, and CW performance at various power levels has been evaluated in a "bar-in-groove" laser package. These high-power 1-cm stacked-bar arrays are now being manufactured for OEM use. Individual diode laser bars, ready for package-mounting by OEM customers, are being sold as commodity items. Commercial and medical applications of these laser arrays include solid-state laser pumping for metal-working, cutting, industrial measurement and control, ranging, wind-shear/atmospheric turbulence detection, X-ray generation, materials surface cleaning, microsurgery, ophthalmology, dermatology, and dental procedures.

INTRODUCTION

Recent advances in semiconductor diode laser technology have led to significant improvements in device power and efficiency (1). These improvements have resulted from the use of extremely thin, optically active epitaxial layers known as quantum wells (2). Such quantum wells (QWs) are grown in our laboratory by metal organic chemical vapor deposition (MOCVD) (3,4). This technology, when properly applied, is capable of yielding epitaxial wafers with excellent compositional, doping, and thickness uniformity over 50 mm and 75 mm diameter GaAs wafers. Continual device improvements are being achieved through a progression of developments which have involved both epitaxial material growth and laser bar packaging techniques.

Recent advances in the use of strained QW epitaxial layers have led to diode lasers with the highest reported values of efficiency and long-term reliability (5). Reliability improvements in these strained layer structures are thought to result from the use of indium in the AlGaAs QW layers; indium which has the ability to stop the propagation of dislocation damage at the QW-barrier layer interfaces (6,7).

For diode laser array bar packaging technologies, two recent developments are relevant. A group at the Lawrence Livermore National Laboratory (LLNL) has designed a diode laser package having extremely low thermal-resistance, utilizing the microchannel cooling concept (8,9). This package is superior for high duty factor diode laser array operation, and has been shown capable of enabling standard 1-cm diode laser bars to emit over 20 watts of CW power with good reliability (10). For shorter duty factor operation (below 20%) and

pulse lengths of the order of 2 ms or less, a "bar-in-groove" package utilizing a metallized ceramic mounted on a copper heat sink has been shown to produce multi-bar diode laser arrays emitting over 1 kW of peak power for 300 μ sec pulses with excellent reliability (11). Larger duty factors (to CW) and longer pulse lengths can be used with high reliability, provided proper laser power derating is used. In this paper we describe recent high power diode laser array performance results for 1-cm wide laser bars mounted in the bar-in-groove package.

HIGH POWER LASER DESIGN AND FABRICATION

The basic epitaxial wafer structure used for the high power diode lasers described here consists of a GRaded INDEX Separate Confinement Heterostructure (GRINSCH) which has previously been shown to give excellent performance (1). There are two techniques by which such epitaxial layers, with active QW regions as thin as 50Å, can be grown -- molecular beam epitaxy (MBE) and MOCVD. Whereas MBE involves ultra-high vacuum deposition equipment and has demonstrated good laser performance, the highest power diode lasers reported to date have been grown by MOCVD. It is the latter technique that was used to produce the high power lasers reported here. Spire has employed MOCVD over the past ten years and has pioneered much of the technology through the development and sale of such epitaxial reactors to the semiconductor community. This reactor development activity is continuing, and most of the epitaxial wafers reported here were grown in a Spire MOCVD-100S, single wafer, low pressure reactor. A block diagram of this reactor is shown in Figure 1. A larger version of this reactor, known as the MOCVD-400S, is presently under construction, and will grow up to 4-inch epitaxial GaAs wafers.

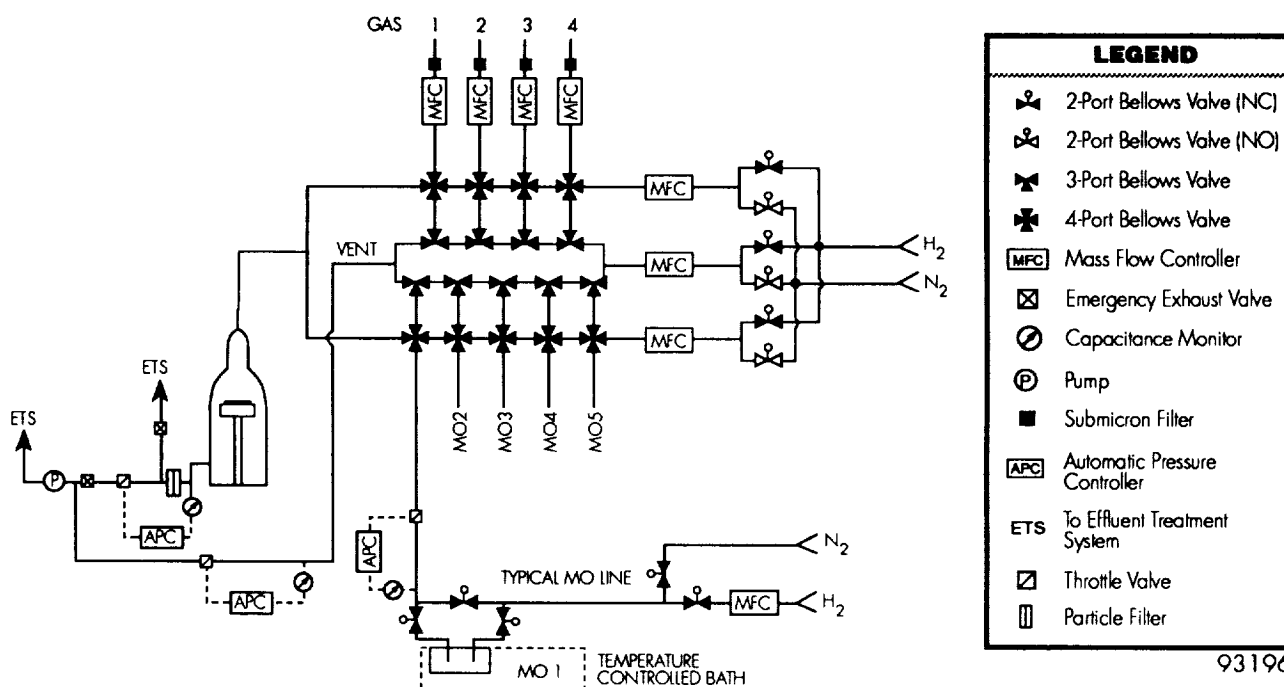


Figure 1 Block diagram of the Spire MOCVD-100S low pressure epitaxial reactor used for the growth of highly uniform epitaxial films.

The basic chemical reaction resulting in the deposition of the epitaxial films involves the pyrolysis of metal alkyls (trimethylgallium and trimethylaluminum) and a hydride (arsine) at temperatures of approximately 700°C in the presence of hydrogen and certain dopant gasses such as silane and dimethylzinc. Excellent film uniformity has been observed for wafers grown in this reactor.

The GRINSCH epitaxial structure, starting with an n-type GaAs substrate with low etch-pit density values, consists of an n-type GaAs buffer layer, an n-type $\text{Al}_{0.6}\text{Ga}_{0.4}\text{As}$ clad layer, the undoped GRINSCH layer with a single 80Å QW and a total thickness of 3000Å, a p-type AlGaAs clad layer of the same composition as the n-clad layer, and a p+ top contact layer for low electrical contact resistance. All layers are grown in one growth operation using a microprocessor driven controller.

Lateral mode control is achieved by gain-guiding in a coupled-mode configuration resulting from 5 μm active regions on 10 μm centers. For quasi-CW pulsed array bars the optical fill factor is over 90%; the entire bar emits, except for 40 μm grooved regions between the active phase-locked emitting regions. A top (p-contact) view of a typical quasi-CW bar 1-cm wide with a 0.5 mm cavity length is shown in Figure 2.

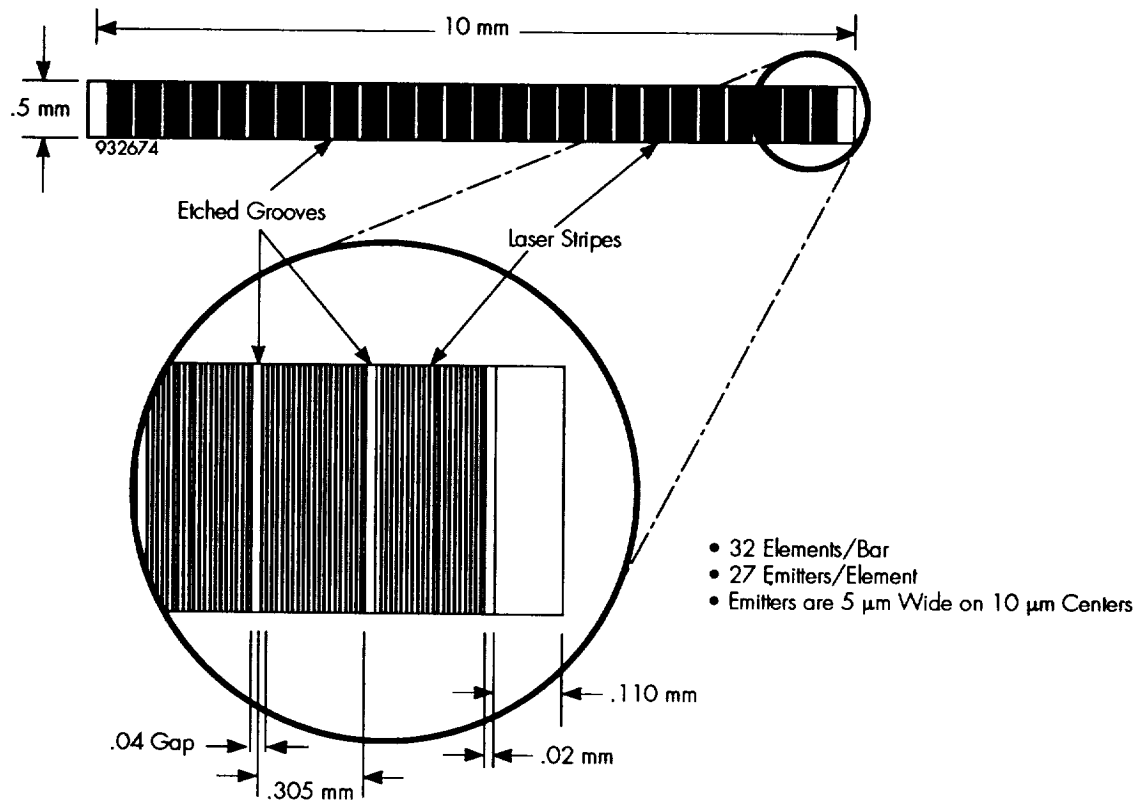


Figure 2 Schematic of the top of a 1-cm wide pulsed diode laser array bar with a 0.5 mm long cavity.

For CW bars, a lower fill factor configuration is used. Such a configuration facilitates more efficient heat removal in normal packages. To enable convenient coupling into optical fibers, the CW bar design consists of twelve sections of active regions, each of which consists of twenty, 5 μm wide active regions on 10 μm centers. The 200 μm wide active regions are on 800 μm centers, giving a 25% optical fill factor. A schematic of the CW array bar configuration is shown in Figure 3.

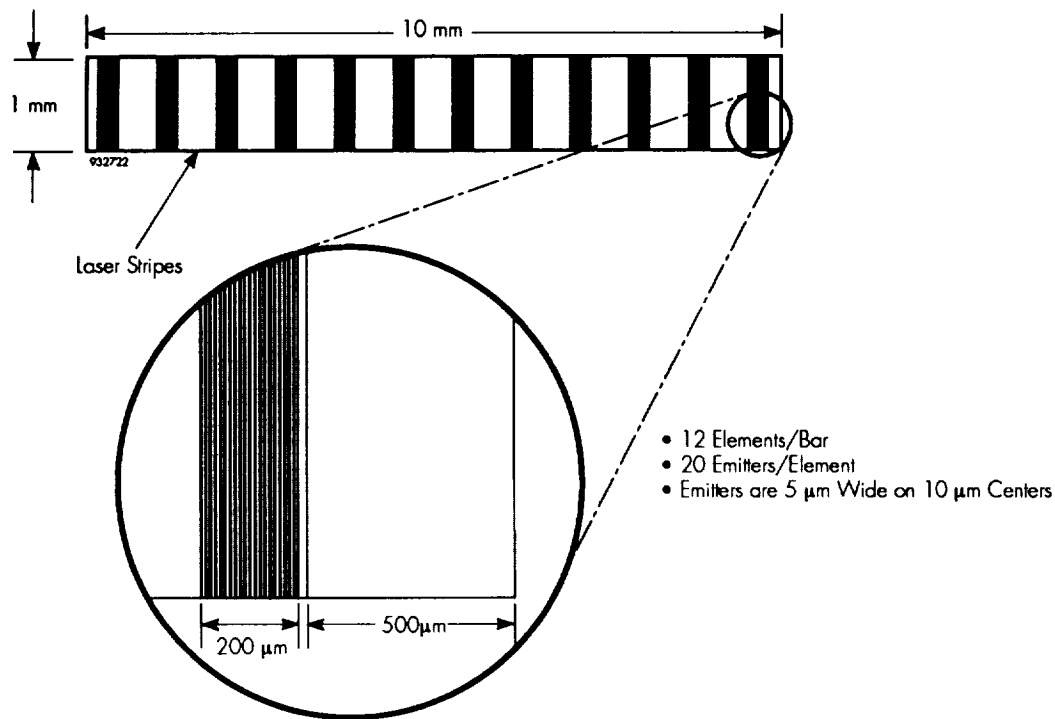


Figure 3 Schematic of the top of a 1-cm wide CW diode laser array bar with a 1 mm long cavity.

The 1-cm cavity length will increase the laser threshold current, but significant improvement in heat sink efficiency is expected.

The gain guiding discussed above is achieved by use of a deposited and patterned silicon nitride or silicon dioxide insulator layer into which the stripe openings are etched. Metallization of the 120 μm thick wafers is carried out on both the n-side and p-sides. In each case, a multi-layer metal system is deposited to ensure metallurgical stability when the laser bars are exposed to high temperatures. Spire's metallization configuration is compatible with the solder required for bar-in-groove mounting. Post-metallization sintering is used to ensure low contact resistance and metallurgical stability.

Individual diode laser array bars are produced by cleaving. Automated equipment provides reproducibility in laser cavity length. Mask codes allow for identification of individual laser bars with their location on each epitaxial wafer. Front and rear facet coatings are applied immediately after cleaving. Rear coatings consist of multilayer dielectric layers with reflectivities of nearly 100%, while the front coatings are half-wavelength with approximately 30% reflectivity. Front coating reflectivity will affect threshold current and output power. After coating, all bars are 100% inspected for cosmetics, particulates and metallization/coating appearance. Any bars which exhibit flaws or defects which could affect operation under high power conditions are rejected prior to assembly. The individual bars are stocked by emission wavelength in gelpak packages for sale to the semiconductor diode laser industry as off-the-shelf, inventory items. Facet colors allow users to distinguish between the front and the rear facets. Representative bars from each wafer are tested for P-I, V-I, efficiency-I, emission spectrum and reliability for 250,000 shots at full operating current. A map of the emission wavelengths measured on laser bars fabricated from a typical 50 mm epitaxial wafer is shown in Figure 4.

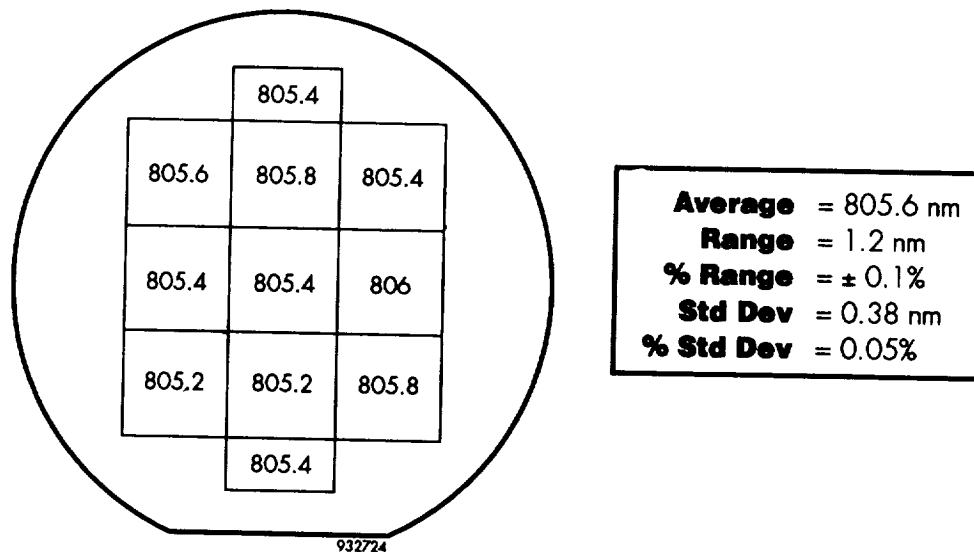


Figure 4 Laser emission wavelength map showing emission wavelength uniformity over a 50 mm diameter epitaxial QW wafer.

The laser bar emission wavelength uniformity has a range of 1.2 nm, with a standard deviation of 0.38 nm. This represents a $\pm 0.1\%$ emission wavelength uniformity, with a standard deviation of only 0.05%.

The overall diode laser array bar fabrication process is summarized in Figure 5.

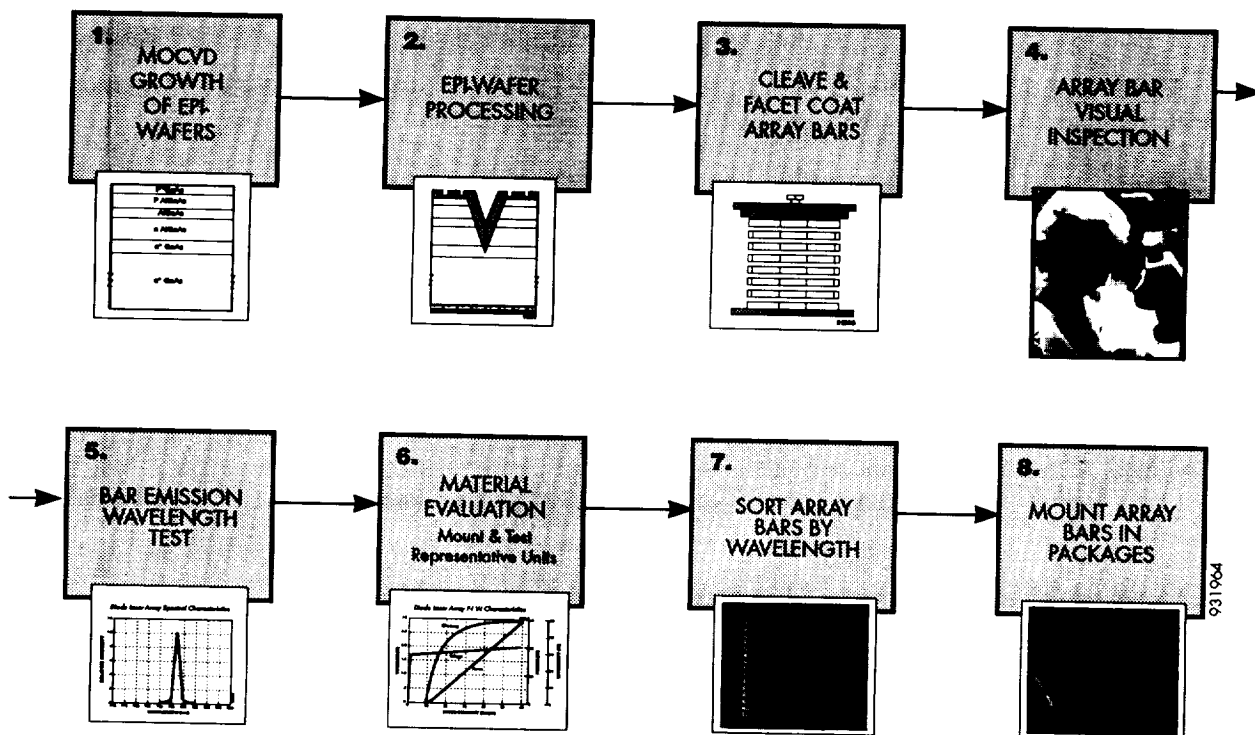


Figure 5 Diode laser array bar manufacturing sequence illustrating the various processing and evaluation steps.

A multiplicity of diode laser array packages are available, depending on the required output power and duty cycle. A photograph of a typical 6-bar diode laser array, with bars mounted in the "Z-format" package by Laser Diode Arrays, Inc. (LDAI), is shown in Figure 6.

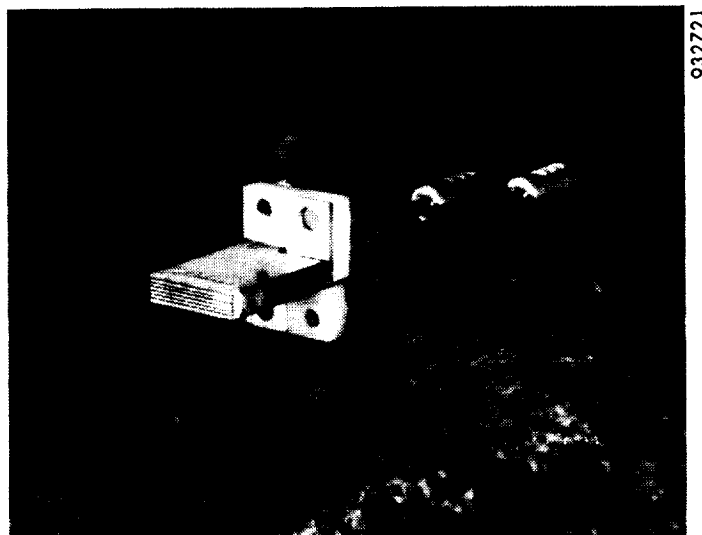


Figure 6 Photograph of the Z-format diode laser array package with 6 emitting bars. Such an array is rated at 300 W, quasi-CW emission (50 W/bar).

Packages of this type with up to 40 bars can be readily fabricated. Some of the critical performance parameters of these lasers and laser arrays are summarized in the next section.

LASER PERFORMANCE DATA

In this section we will present some of the important performance parameters of multibar diode laser arrays for quasi-CW operating conditions. With typical threshold current values of 8 A, these devices exhibit electrical-to-optical conversion efficiency values of from 45 to 50%. Figure 7 shows the P-I, V-I and efficiency-I characteristics of a typical 6-bar array of the type shown in Figure 6.

Each array is burned-in for at least 250,000 shots prior to delivery. No changes in performance are observed during this burn-in period. The array of Figure 7 was burned-in with 1 ms pulse lengths at a 10 Hz repetition rate. With bars on 0.5 mm centers, the array exhibited a power density of 1240 watts/cm². Arrays of this type have been operated at 100 million shots with a measured reduction in output power of approximately 2%.

A typical emission spectrum for this type of 6-bar array is illustrated in Figure 8.

The three curves shown in this figure represent emission spectra for 100, 250 and 500 μ sec long pulses. As can be seen, there is not much change in emission spectrum with pulse length under these conditions. The operating current of this array is 50 A at room temperature. The FWHM of the emission spectrum is approximately 2.5 nm. Typical far-field patterns (full width, half maximum, FWHM) are 38 degrees (fast axis, perpendicular to the junction plane) by 10 degrees (slow axis, along the junction plane). The polarization (E-vector) is along the junction direction.

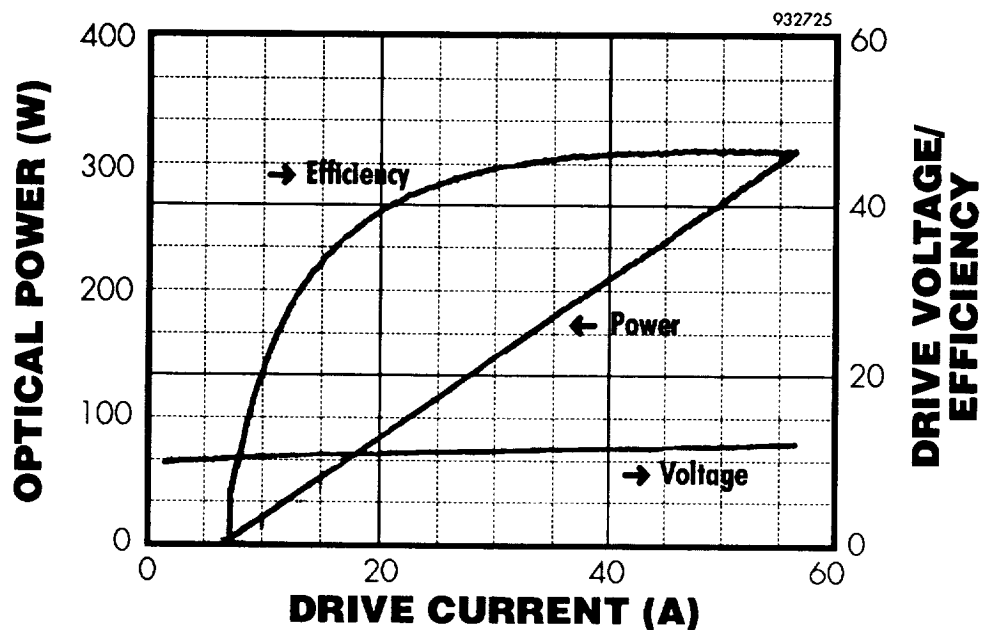


Figure 7 P-I, V-I and efficiency-I characteristics of a typical 6-bar diode laser array of the type shown in Figure 6.

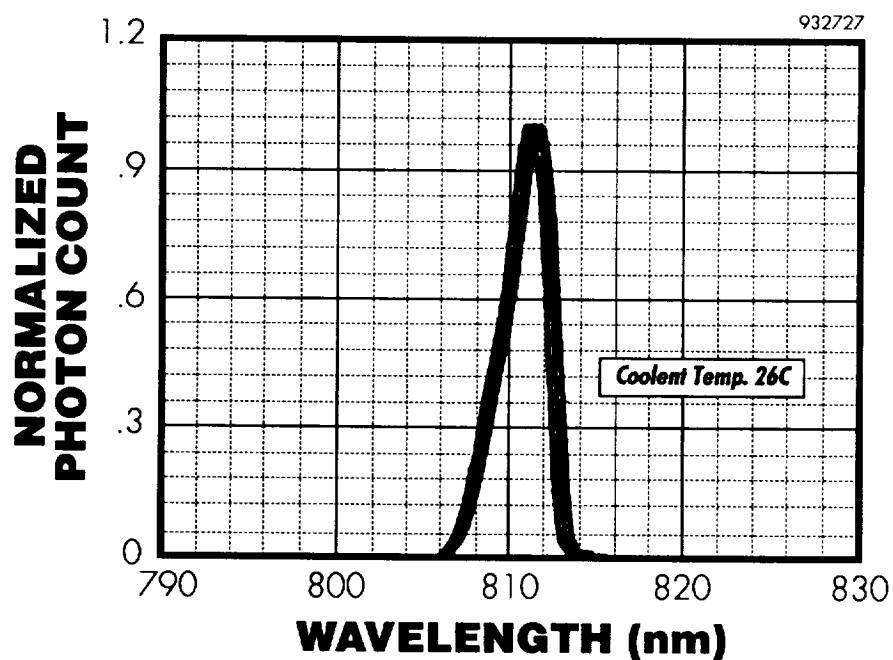


Figure 8 Emission spectrum of a typical 6-bar, 810 nm, bar-in-groove, quasi-CW diode laser array under various pulse-width conditions.

The performance of typical 6-bar diode arrays in the bar-in-groove package has been measured under conditions of heavy thermal stress by the NASA Langley Research Center. Figure 9 shows the measured variation in FWHM of the emission spectrum at a variety of very long pulses (up to 3.5 milliseconds).

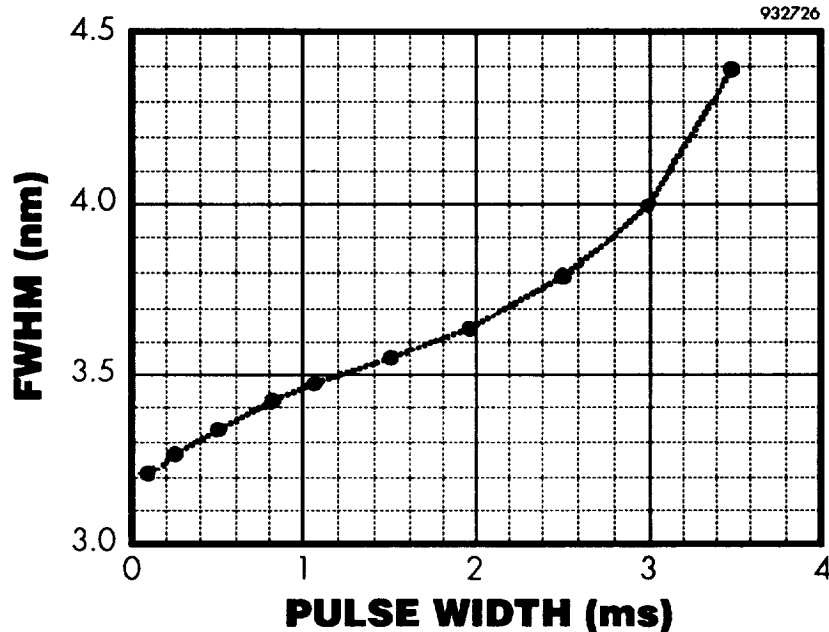


Figure 9 Measured FWHM of the emission spectrum of a typical 6-bar, quasi-CW diode laser array mounted in the LDAI bar-in-groove package.

The increase in FWHM is a result of the frequency chirping that occurs during the long on-pulses, and is caused by thermal heating of the array bars. We estimate the thermal time constant (ratio of heat capacity to thermal conductivity per unit area) of this type of package to be in the 40 to 60 ms range for 1-cm bars with a 0.5 mm cavity length. It is thus important to keep the time-separation between pulses well above this range if pulse lengths of greater than 2 or 3 ms are used.

A graph of the center wavelength of the emission peak as a function of pulse width for this type of array is shown in Figure 10. With a pulse repetition rate of 10 Hz, the duty cycle is 1 % for a 1 ms pulse and increases to 5% for a 5 ms width. Increasing pulse width increases heating of the bars, thereby shifting the emission wavelength towards long wavelengths. The emission wavelengths shown in this graph represent the average observed over the pulse lengths indicated. During the actual pulses, chirping occurs, as represented by the increasing FWHM values of Figure 9. With a known wavelength shift of approximately 0.25 nm/degree C, we estimate an average junction temperature increase of 4 degrees for every ms increase in pulse length. The actual junction temperature at the end of each of these long pulses can be considerably larger than this.

The near-field emission pattern of diode laser arrays mounted in the bar-in-groove configuration can be observed with either a CCD camera (home video camera) or with ordinary color film using a 35 mm camera. Figure 11 shows a head-on view of a 15-bar diode laser array. Part (a) shows the cut grooves with bars inserted, while part (b) is a 35 mm photograph taken of the array operating just above threshold. As can be seen, every portion of each 1-cm bar is lit up, with no dead spots.

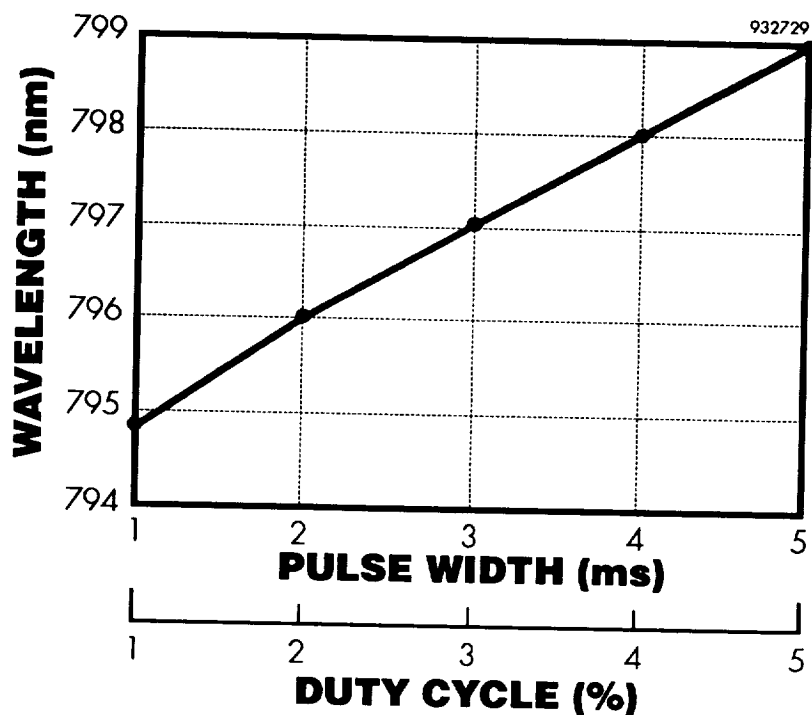


Figure 10 Emission wavelength vs. pulse width (10 Hz drive frequency) for a 6-bar diode laser array.



a)



b)

Figure 11 Head-on photographs of a 15-bar diode laser array consisting of 15 one-cm bars mounted in narrow grooves cut into a metallized, thermally-conducting ceramic plate. Part (a) shows the grooves with the laser bars inserted, while part (b) shows the light emitting regions just above threshold, as seen on a 35 mm film.

The performance characteristics of a 32-bar array fabricated from four sections of 8-elements each, stacked side by side to provide an emitting region of 4-cm by approximately 0.5 cm are shown in Figure 12. This array emitted over 1.2 KW of peak power with 250 μ s pulses at a 10 Hz repetition rate. The FWHM of the emission wavelength of the entire array was measured at 4 nm for 100 μ s pulses, and increased to only 4.5 nm for 500 μ s pulses. The array is presently in use in a commercial laser system, and has shown outstanding performance reliability.

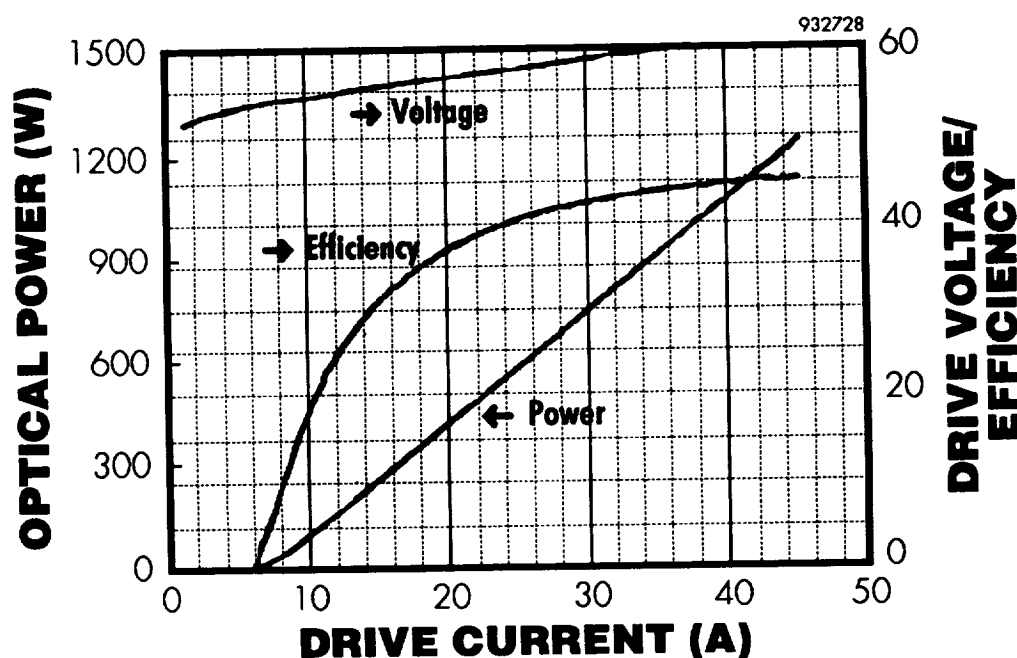


Figure 12 P-I, V-I and efficiency-I characteristics of a 32-bar diode laser array emitting over 1.2 KW of peak power with 250 μ s pulses at 10 Hz.

LASER RELIABILITY

Long-term reliability is one of the most important issues to be addressed if wide-spread use of high power diode laser arrays is to be achieved. Such reliability is effected by the quality of the GaAs substrate wafers used, the structure and quality of the epitaxial layers, the metallization technique used for both the n- and the p-side of the epitaxial wafers, the facet coatings, and the handling of the epitaxial wafers and finished bars. Experience has shown that with proper manufacturing and handling techniques, excellent reliability can be achieved. Spire laser array bars have been evaluated in a variety of packages, and under various operating conditions. The longest pulsed-performance data available to date has been obtained by Jenoptik GmbH in Wiesbaden, Germany, where a 10 billion shot test is in progress. To date, almost 2 billion shots have been carried out on a collection of 6 1-cm bars. The data are shown in Figure 13.

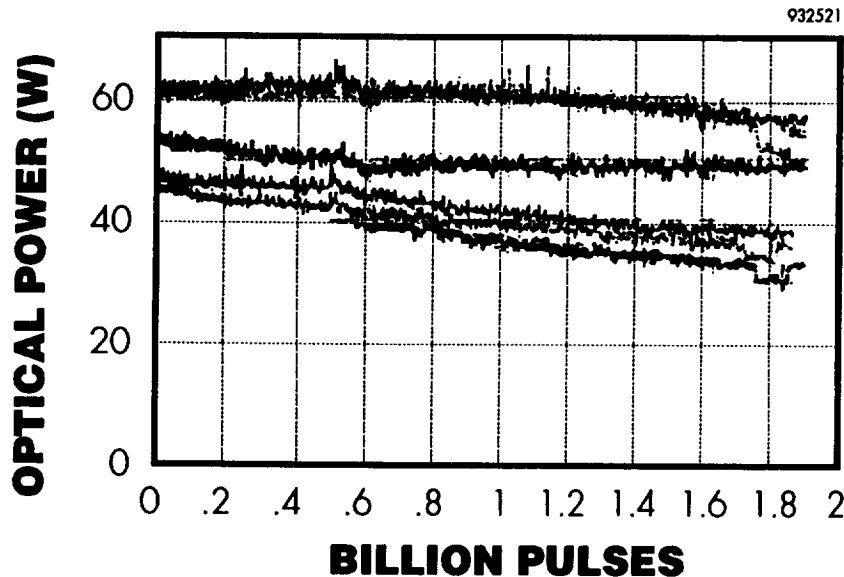


Figure 13 Long-term reliability data for Spire high power diode laser bars (obtained by Jenoptik GmbH). The bar tests are currently at over 2 billion shots, and are scheduled for termination at 10 billion shots.

The tests are being conducted at room temperature, with 300 μ s pulses, 60 A current, and 135 Hz repetition rate (4% duty factor). These data are for a random sampling of laser bars, but all bars had passed a critical visual and electrical inspection prior to mounting.

Additional reliability data have been obtained at other laboratories. Lawrence Livermore National Laboratory has mounted two bars (reduced in length to 0.9 mm by cleaving the ends off) in one of their microchannel coolers (8). In spite of the excessive mis-handling (chopped off ends) of these bars, they have operated at 22 W per bar (44 W, CW, for the two bars) for 70 hours with no degradation (10). Additional reliability tests of Spire materials are presently in progress at this laboratory.

FUTURE AREAS OF DEVELOPMENT

Recent developments have indicated that the operation of high power diode laser arrays can be extended both into the visible spectral region and further into the infrared. High power visible diode lasers emitting in the red (633 nm) region of the spectrum have been reported (12). The GaInP/GaInAlP/GaAs visible materials have been extensively studied in Japan and are becoming widely available from companies such as NEC, Sony, Toshiba, Sumitomo Electric, Matsushita Electric and Hitachi, mostly in low power versions, for printing and bar-code readers. High power diode lasers of this material system are currently under study at a number of laboratories, and Spire also is involved in their development. Such short wavelength lasers have potential applications in pumping the tunable solid-state lasers, LiSAF, LiCAF, and Alexandrite, whose optical absorption regions tend to be quite broad. They also have medical applications for ophthalmology and photodynamic therapy (PDT).

In the longer wavelength infrared, the recent announcement of 2-5 μ m diode lasers fabricated from GaInAsSb/AlGaAsSb represents a major advance in the state of the art (13,14). Prior to this announcement, the only diode lasers operating in this spectral region were those fabricated from lead-chalcogenide materials, and

such lasers require cryogenic cooling. Recently, near IR lasers emitting in the 2 μm region have been fabricated from the InGaAsP material system (15). This system is somewhat simpler to handle by MOCVD than the Sb-based compounds (which were grown by MBE at the MIT Lincoln Laboratory), and Spire is presently involved with their development as well. Such high power lasers can be used for pumping Ho:YAG solid state lasers at 1.9 μm , with significantly less quantum heating than that encountered in conventional 792 nm pumping of the Ho,Tm:YAG and YLF solid-state laser systems.

SUMMARY AND CONCLUSIONS

We have described recent progress resulting from a development effort supported, in part, by the NASA Langley Research Center through two Phase I and two Phase II SBIR programs aimed at developing a technology for producing low-cost, high power diode laser arrays emitting in the 800 nm spectral region for solid-state laser pumping. Such high power diode lasers are used as efficient and reliable optical pumps for the Nd:YAG, Nd:YLF and Ho,Tm:YAG solid-state laser systems. The technical achievements of these programs, together with continued IR&D support and a strong corporate commitment, have positioned Spire as a leading supplier of commercial epitaxial laser wafers, high power laser array bars, and high power diode laser arrays. Spire now offers such products, some as off-the-shelf items, to the semiconductor laser industry and to the optoelectronic OEM community. These products are being sold domestically and overseas. Spire's standard products include epitaxial wafers of the GRINSCH-SQW or multi-QW design, 50 W diode laser array bars in standard 1-cm wide configurations, metallized and ready for mounting, as well as multi-bar diode laser arrays with 45 - 50% efficiency and reliability values exceeding 2 billion shots. Such bars have been operated at up to 100 W of pulsed output power, and with pulse lengths of 5 milliseconds. Reliability evaluation trials of 300 μs pulsed devices will be terminated at 10 billion shots. With good high-duty factor heat sinks, Spire's standard pulsed laser designs have been operated under continuous, CW conditions. Lower fill-factor CW laser designs are presently nearing the final phase of development and will be introduced shortly. The company is actively involved with expanding the wavelengths of operation of such arrays, both towards the visible and the infrared regions, thereby opening more areas of application.

ACKNOWLEDGMENTS

Much of the work described here was funded by the NASA Langley Research Center through SBIR Phase II Contracts NAS1-18660 and NAS1-19301; the author is most grateful to Dr. C. J. Magee for his support, encouragement and technical contributions. The author is also indebted to valuable technical input on solid-state lasers by Dr. Norman Barnes of NASA. Materials and device fabrication and testing by Edward Gagnon, Leo Geoffroy, Victor Haven, Andre Mastrovito and Michael Sanfacon of the Spire Corporation are also acknowledged. Multibar packaging of Spire's diode lasers has been carried out by Laser Diode Arrays, Inc., and the author acknowledges the contributions to this endeavor by Mr. Alan Karpinski. We also acknowledge the contribution by Jenoptic GmbH in generating long-term reliability data for our high power laser array bars.

REFERENCES

1. W. Streifer, D. R. Scifres, G. L. Harnagel, D. F. Welch, J. Berger and M. Sakamoto, "Advances in Diode Laser Pumps", IEEE Jour. Quant. Elec. 24, 883 (1988).
2. Y. Arakawa and A. Yariv, "Quantum Well Lasers - Gain, Spectra, Dynamics", IEEE Jour. Quant. Elec. 22, 1887 (1986).
3. G. B. Stringfellow, Organometallic Vapor-Phase Epitaxy, Academic Press (1989).
4. K. J. Linden, "Epitaxial Wafers: Semiconductors for Photonics", Photonics Handbook, Laurin Publishing Co., Pittsfield, MA (1993).
5. H. Morkoc, B. Sverdlov and G. B. Gao, "Strained Layer Heterostructures, and their Applications to MODFETs, HBTs, and Lasers", Proc. IEEE 81, 493 (1993).

6. R. G. Waters, R. J. Dalby, J. A. Baumann, J. L. De Sanctis and A. H. Shepard, "Dark-Line-Resistant Diode Laser at 0.8 μm Comprising InAlGaAs Strained Quantum Well", *IEEE Photonics Technol. Lett.* 3, 409 (1991).
7. S. L. Yellen, A. H. Shepard, R. J. Dalby, J. A. Baumann, H. B. Serreze, T. S. Guido, R. Soltz, K. J. Bystrom, C. M. Harding and R. G. Waters, "Reliability of GaAs-Based Semiconductor Diode Lasers: 0.6 - 1.1 μm ", *IEEE Jour. Quant. Elec.* 29, 2058 (1993).
8. R. Beach, W. J. Bennett, B. L. Freitas, D. Munding, B. J. Comaskey, R. W. Solarz and M. A. Emanuel, "Modular Microchannel Cooled Heatsinks for High Average Power Laser Diode Arrays", *IEEE Jour. Quant. Elec.* 28, 966 (1992).
9. W. J. Bennett, B. L. Freitas, D. Ciarlo, R. Beach, S. Sutton, M. Emanuel and R. Solarz, "Microchannel cooled heatsinks for high average power laser diode arrays", *Proc. SPIE* 1865, 144 (1993).
10. R. Beach, private communication.
11. A. Karpinski and K. J. Linden, "Novel Packaging for High-Performance Low-Cost Diode Laser Arrays", *OSA Tech. Digest* 16 (Nov., 1990).
12. D. P. Bour, D. W. Treat, R. L. Thornton, T. L. Paoli, R. D. Bringans, B. S. Krusor, R. S. Geels, D. F. Welch and T. Y. Wang, "Low threshold 633 nm, single tensile-strained quantum well GaInP/AlGaInP laser", *Appl. Phys. Lett.* 60, 1927 (1992).
13. S. J. Eglash and H. K. Choi, "Efficient GaInAsSb/AlGaAsSb Diode Lasers Emitting at 2.29 μm ", *Appl. Phys. Lett.* 57, 1292 (1990).
14. H. K. Choi and S. J. Eglash, "High-power multiple-quantum-well GaInAsSb/AlGaAsSb diode lasers emitting at 2.1 μm with low threshold current density", *Appl. Phys. Lett.* 61, 1154 (1992).
15. S. Major, D. Nam, J. Osinki and D. Welch, "High power 2.0 μm InGaAsP laser diodes", *IEEE Photonics Technol. Lett.* 5, 594 (1993).

509-31
2512
p-6

FLEXIBLE MANUFACTURING FOR PHOTONICS DEVICE ASSEMBLY

Shin-ye Lu
Thrust Area Leader, Engineering Research Division, Engineering Department
Michael D. Pocha
Group Leader, Engineering Research Division, Engineering Department
Oliver T. Strand
Physicist, L-Division, Nuclear Test Experimental Sciences Department
K. David Young
Project Engineer, Engineering Research Division, Engineering Department
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94550

The assembly of photonics devices such as laser diodes, optical modulators, and opto-electronics multi-chip modules (OEMCM), usually requires the placement of micron size devices such as laser diodes, and sub-micron precision attachment between optical fibers and diodes or waveguide modulators (usually referred to as pigtailling). This is a very labor intensive process. Studies done by the opto-electronics (OE) industry have shown that 95% of the cost of a pigtailed photonic device is due to the use of manual alignment and bonding techniques, which is the current practice in industry. At Lawrence Livermore National Laboratory, we are working to reduce the cost of packaging OE devices through the use of automation. Our efforts are concentrated on several areas that are directly related to an automated process. This paper will focus on our progress in two of those areas, in particular, an automated fiber pigtailling machine and silicon micro-bench technology compatible with an automated process.

INTRODUCTION

At present, the cost of opto-electronic devices is dominated by the effort required to package those devices into an integrated system. Components such as laser diodes and modulators, designed for high-performance applications, are single-mode devices; they must be connected together using optical fibers or other type of waveguide with sub-micron alignment accuracies. Presently, the OE packaging is usually performed by highly skilled technicians looking through microscopes and manually adjusting sub-micron stages. For single mode fibers, six degrees of freedom for positioning are sometimes required. Once the alignment is correct, the components must be held in place using epoxy, solder, or other attachment techniques, and realigned before the gluing process is settled. This labor-intensive process results in only a few packages being produced per day by each technician. The packaging costs are by far the highest fraction of the total cost of an assembled OE package. The consequences of this low-volume labor-intensive process of packaging OE devices are readily apparent. The costs are too high to allow the advantages of fiber optics to penetrate such markets as on-chip interconnects, interboard connections in computers, and local area networks.

At LLNL, we believe that the packaging process must be automated to significantly reduce the costs of OE devices. The electronics industry has successfully reduced the costs of its products through the massive use of automation, including alignment, parts handling and feeding, and in-situ quality control. A simple model (Figure 1), which takes into account the initial cost of the automated machinery, the labor costs of an operator, and the material costs of the devices, shows that substantial cost savings may also be realized in the opto-electronics industry at even modest production rates. Unfortunately, the sub-micron precisions, and six-axis alignment required for OE packaging greatly exceeds the requirements of the electronics industry. The automated systems developed to assemble integrated circuits cannot be applied to the problem of packaging opto-electronic circuits.

Work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

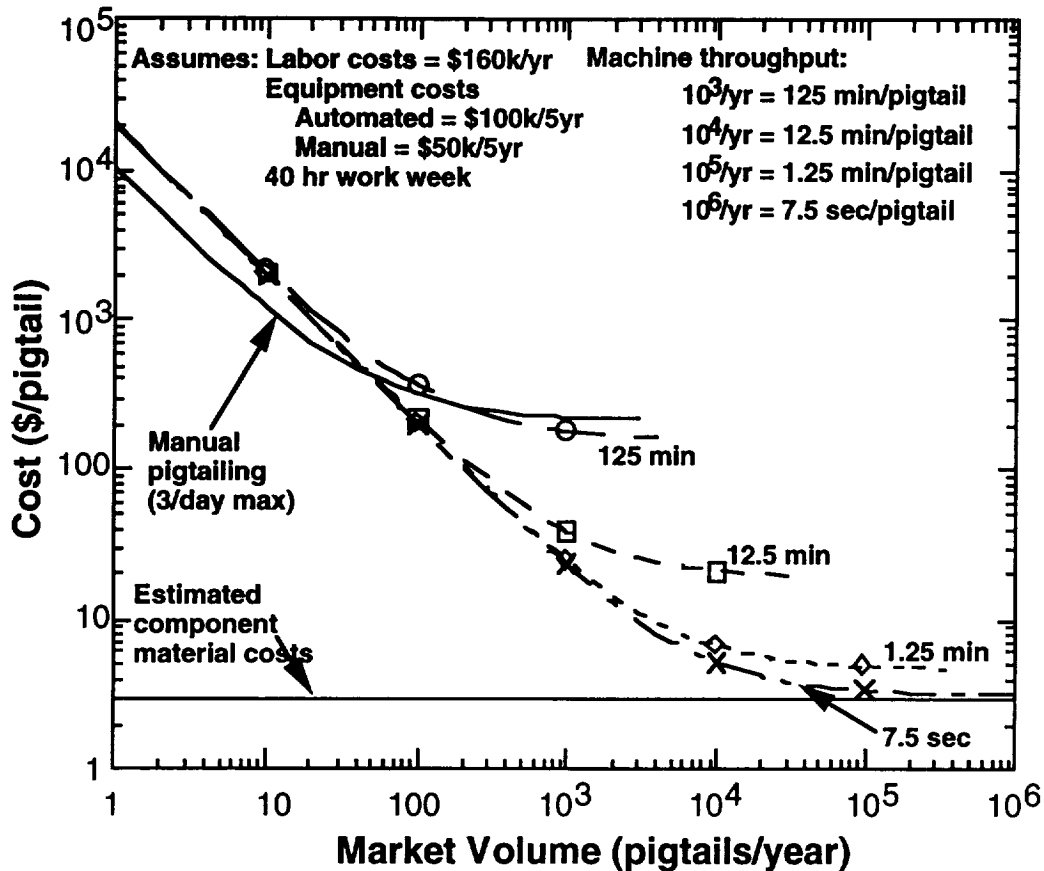


Figure 1. Cost Analysis of an Automated Opto-electronic Packaging Machine

We are currently designing and building a flexible workcell for automating the packaging of OE devices (Figure 2). The workcell design uses machine vision to provide position information for low resolution (microns) long travel movement, and optic power throughput optimization to provide final high resolution small travel alignment. The combined vision/optic power feedback approach provides a low cost output metrology for high resolution positioners. It also eliminates the need of high precision fixturing, calibration, and temperature control to hold photonics devices in fixed locations at micron precision. The workcell should be able to handle a wide variety of different sizes and attachment geometry of photonic devices.

The initial task is to align a single-mode fiber to each end of a Mach-Zehnder waveguide modulator (fiber pigtail). We are also developing silicon micro-bench based techniques to provide a packaging technology that is compatible with automated alignment and pigtail.

MACHINE VISION SYSTEM

There are several parameters that need to be considered in the design of a system for machine vision: pixel resolution, field of view (FoV) depth of field, and the cost. Ideally, the vision system should have high pixel resolution to yield position accuracy, large field of view to cover the entire size and depth range of the photonic device, and low cost. We decided to use a low cost, standard microscopic camera commonly used for industrial inspection. The image frame grabber produces 640-by-480 pixels in one view. The pixel resolution is 1.44 micron. The active view volume is approximately 900 x 700 x 100 cubic microns, which can be substantially smaller than a photonic device to be assembled. Currently, the automated packaging workcell consists of two microscopic camera systems that provide a top view and a side view of the substrate and the fibers. Each camera is mounted on motorized stages to cover the entire area of interest.

The motion control hardware and software for the vision system are chosen for position accuracy to be compatible with the image resolution such that the static position repeatability (without the use of vision feedback) of the vision system is 2 to 5 microns. There are two reasons for the high position repeatability requirement of the

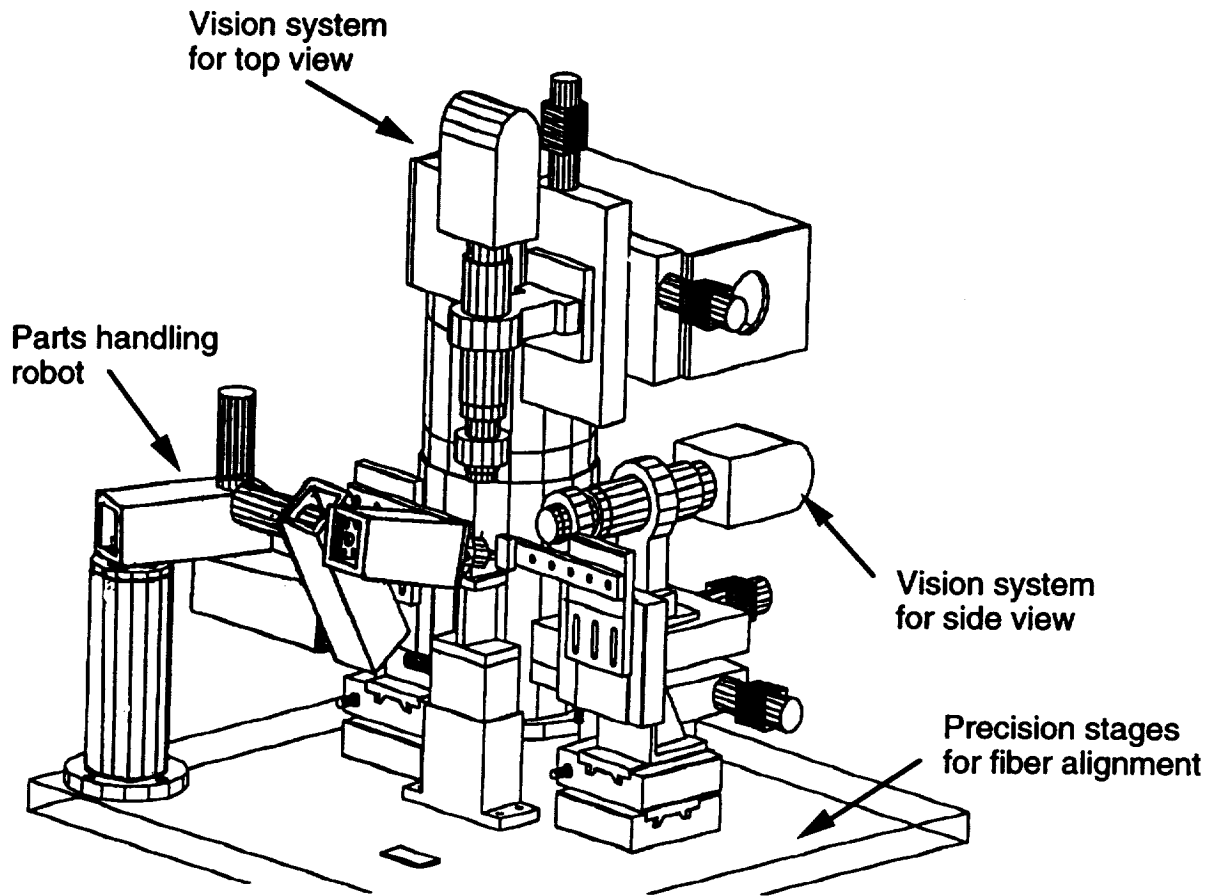


Figure 2. LLNL Automated Opto-electronics Packaging Machine

vision system: (1) to move from one FoV to the next along a certain trajectory, (2) to move between two focal planes, and be able to register the two coordinate systems defined by the two images.

The top camera system has three-degrees of freedom that are configured as a scalar arm to move around the horizontal plane -- one rotary motor swings the arm and one linear stage slides along the arm. The camera is mounted on another linear slide perpendicular to the arm to move along the optic axis. The rotary motor used in the top camera system has 600,000 counts per revolution. The rotary arm is ten inches long. Both sides have two inches travel and one micron repeatability. The repeatability of the top camera system is approximately $2.5 \times 1 \times 1$ cubic microns. The side camera has two-degrees of freedom for moving along a horizontal line and for moving along the optic axis. The same type of slides are used for the side camera system.

VISION FEEDBACK FOR KINEMATIC CALIBRATION

The static position repeatability is a necessary but not a sufficient condition to register the coordinate systems defined by two FoV of two focal planes. The misalignment of the mechanic system also needs to be calibrated to generate the necessary parameters for coordinate transformation.^{1,2} Since the side camera represents a subsystem of the top camera system (without the rotational degree of freedom), we will describe the calibration procedure using vision feedback to back-calculate the kinematic configuration of the top camera system only.

The calibration procedure utilizes a sequence of predetermined motions of the top camera system. Stationary fiducials on the substrate are tracked by their respective pixel coordinates in the sequence of focal plane images. Of the many kinematic parameters that can be calibrated, we concentrate on the following three: The misalignment angle of the Camera Frame with respect to Linear Stage Frame, δ , and the coordinates of the origin of the Camera Frame as measured in the Linear Stage Frame coordinate system, h and l . The orientation of the Linear Stage Frame is changed by a motor drive system through the control of the angular displacement, $\Delta\theta$. Fig. 3

illustrates the kinematic parameters and their relationships to the fiducials, the Camera Frame and Linear Stage Frame.

The misalignment angle is determined by moving the top camera along the linear stage translation axis, as measured by the Δl increments, while keeping the rotation degree of freedom fixed. The linear stage is commanded to move a small distance at a time while keeping the same set of fiducials in the field-of-view of the camera at all times. Selected points of the fiducials are tracked and their pixel coordinates on the focal plane image are registered. The misalignment angle is calculated from the pixel coordinates of a given fiducial point corresponding to its location before (x_1 and y_1) and after a commanded move (x_2 and y_2) by

$$\tan \delta = \frac{x_2 - x_1}{y_2 - y_1} \quad (1)$$

The coordinates of the image plane origin are determined by rotating the top camera while keeping the linear degree of freedom fixed. The motor drive is commanded to move in a small angle increments $\Delta\theta$, while the same set of fiducials are kept within the field-of-view. Again, selected points of the fiducials are tracked and their pixel coordinates on the image are registered. Using the following set of linear equations, the coordinates, h and l , are computed from the pixel coordinates of a given fiducial point corresponding to its location before (x_1 and y_1) and after the move (x_2 and y_2), the previously calculated misalignment angle, δ , and the angle increment, $\Delta\theta$.

$$\begin{aligned} (1 - \cos \Delta\theta)l - \sin \Delta\theta h &= (y_1 - y_2 \cos \Delta\theta - x_2 \sin \Delta\theta) \cos \delta \\ &+ (x_1 - x_2 \cos \Delta\theta + y_2 \sin \Delta\theta) \sin \delta \end{aligned} \quad (2)$$

$$\begin{aligned} \sin \Delta\theta l + (1 - \cos \Delta\theta)h &= (x_1 - x_2 \cos \Delta\theta + y_2 \sin \Delta\theta) \cos \delta \\ &+ (x_2 \sin \Delta\theta - y_1 + y_2 \cos \Delta\theta) \sin \delta \end{aligned} \quad (3)$$

SILICON MICRO-BENCH

At LLNL, we are also working on silicon micro-bench techniques such that all components of an opto-electronic multi-chip module (OEMCM) can be attached to flat substrates with direct access from the top. Thus, the process lends itself to a simple packaging procedure. The substrates consist of silicon with gold metalization pads to which the components are attached and the electrical connections are made. Unique polycrystalline silicon on-board heaters are used to quickly attach by reflowing solder the components and single mode fiber pigtailed to sub-micron positioning tolerances.

Prototype silicon micro-benches were developed to pigtail high-powered 800 nm laser diodes to single mode fibers. These micro-benches are 13 mm long by 6 mm wide and 0.5 mm thick. The success of the prototype has led us to develop several follow-up designs. For example, the micro-bench shown in Figure 4 is for packaging a 1550 nm DFB laser. On the left side of the micro-bench, we photolithographically pattern various gold pads to provide a ground plane for the laser and stress relief pads for the wire bonds. For the fiber attachment on the other side of the micro-bench, we build two heating elements made of polysilicon which are attached to gold bonding pads to provide electrical contact points. In the center of each heater, we pattern a gold pad on a layer of silicon dioxide. This gold pad provides the solder attachment point while the silicon dioxide electrically isolates the gold pad from the polysilicon heater. The gold pads are 1 mm by 0.5 mm each and are sufficiently large to solder a 250 micron diameter fiber at the two attachment points. Presently, we use either 100-micron diameter solder balls or solder paste to attach the metalized fiber.

The performance of the polysilicon heaters on our prototype is very reproducible with a specially constructed power supply that allows us to accurately control the magnitude and time of the applied current. Fiber positioning is done by active alignment to sub-micrometer tolerances. While the fiber is held in the position that achieves maximum optical coupling, solder is reflowed to lock the fiber in place. We typically apply one amp of current for approximately 0.5 secs. to reflow solder at the fiber attachment points. We observe no decrease in the amount of light coupled from an 800 nm laser diode into a single-mode fiber before and after the solder reflow and cooling.

Our micro-bench geometries with on-board heaters allow rapid attachment of not only the fiber, but all other components to be placed on the micro-bench. Applying large currents for longer periods of time allows solder reflow at other locations on the micro-bench. Using solders with different melting temperatures and judiciously choosing the order of attachment allows a variety of components to be soldered to the micro-bench without movement of previously attached components. Generally, components furthest from the heater will be attached first using a high current through the heater. We can solder a thermoelectric cooler, a thermistor, and a laser diode onto our micro-bench at different distances from the heater in less than 15 minutes. The placement of these components does not require sub-micron alignment, and can therefore be aligned using standard techniques used in the electronics industry, which is heavily automated. We envision that the placement and soldering of these components onto the micro-benches could be performed by an automated system in only a few minutes.

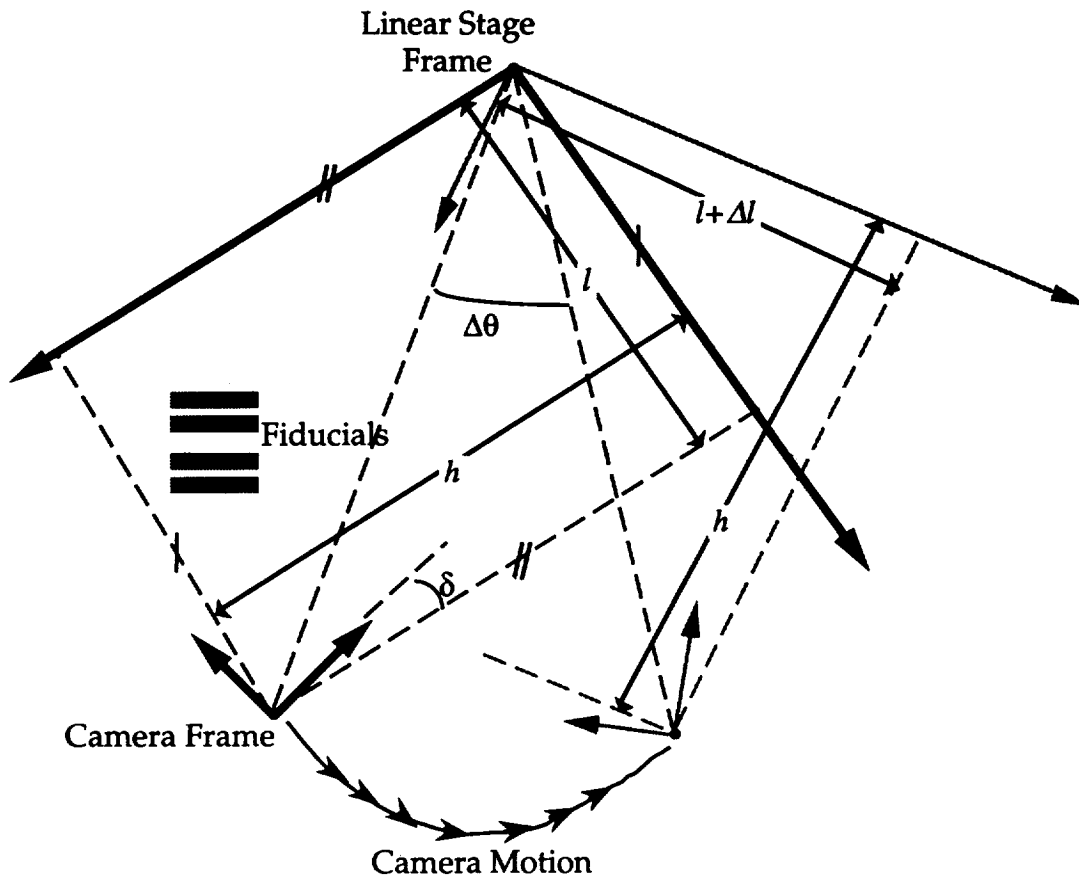


Figure 3. Top Camera System Kinematics

As the last step, the fiber must be aligned to sub-micron tolerances and is attached using the least amount of current through the heaters. The vision system is used to search for those soldered components and attachment points based on fiducial markings. Fiber placement and alignment is guided by the position information generated from computer vision programs to 2 to 5 micron precisions. The final sub-micron precision is provided by an optical power feedback system.³

The idea of on-board heaters lend itself to applications other than packaging laser diodes. We are presently designing a longer micro-bench with heaters at each end to pigtail both ends of a semiconductor optical amplifier. We are also investigating geometries compatible with high speed applications in which on-board transmission lines will be needed to provide sufficient bandwidth.

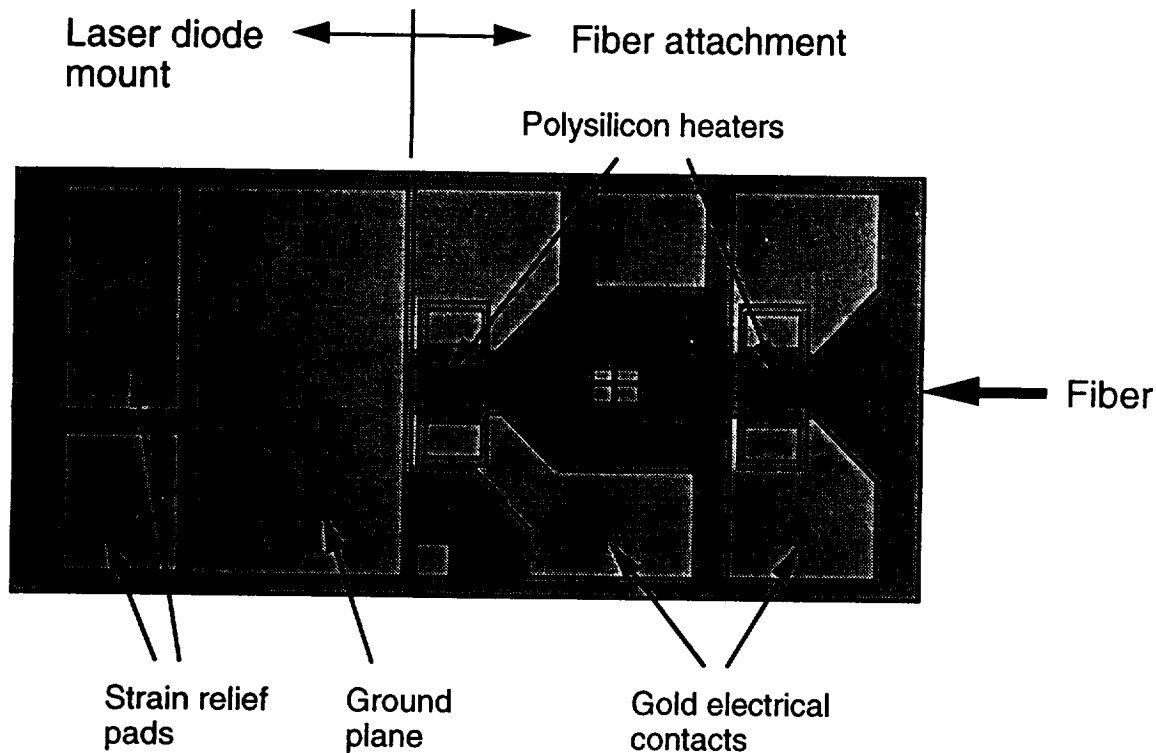


Figure 4. Sketch of our Micro-bench

SUMMARY

The key element in reducing the costs of packaged opto-electronic devices is to minimize the manual labor costs. According to our model, manual pigtailed techniques will not allow the cost of a pigtailed OE device to drop below approximately \$250. On the other hand, an automated process could allow the costs to drop as low as \$10 per pigtail—a factor of 25 decrease in cost. A fully automated system for fiber pigtailed must include automated fiber alignment, fiber attachment techniques which are compatible with an automated process, in-situ quality control, and automated parts handling and feeding. We present here our efforts in addressing the alignment issue and a fiber attachment technique. Variations of an automated system such as this could perform not only sub-micron active alignments as we discuss here, but also passive alignments using tactile feedback rather than optical feedback. The important point is to automate the process regardless of the alignment technique involved. We believe that a massive market is ready for the technology that opto-electronics can provide and is just waiting for the costs to be reduced.

ACKNOWLEDGEMENT

The authors would like to thank Mark Lowry for his guidance and support of the photonics packaging automated effort.

REFERENCES

- (1) "Characterization and Pigtailed of Advanced Waveguide Devices using AutoAlign", Scott Jordan in collaboration with Lawrence Livermore National Laboratory, Newport Corporation Internal Report, April 1993.
- (2) "Determination of Camera Location from 2D to 3D Line and Point Correspondences", Y. Liu, T. S. Huang, O. D. Faugeras, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 1988.
- (3) "Vision-based Automatic Theodolite for Robot Calibration", M. R. Driels, U. S. Pathre, IEE Trans. on Robotics and Automation, Vol. 7, No. 3, June 1991.

0011

ROBOTICS

**APPLICATION OF DEXTEROUS SPACE ROBOTICS TECHNOLOGY
TO
MYOELECTRIC PROSTHESES**

**Clifford Hess and Larry C. H. Li
Automation and Robotic Division
NASA Johnson Space Center
Houston, TX 77058**

**Kristin A. Farry and Ian D. Walker
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77251**

530-37

2513

p. 14

ABSTRACT

Future space missions will require robots equipped with highly dexterous robotic hands to perform a variety of tasks. A major technical challenge in making this possible is an improvement in the way these dexterous robotic hands are remotely controlled or teleoperated. NASA is currently investigating the feasibility of using myoelectric signals to teleoperate a dexterous robotic hand. In theory, myoelectric control of robotic hands will require little or no mechanical parts and will greatly reduce the bulk and weight usually found in dexterous robotic hand control devices. An improvement in myoelectric control of multifinger hands will also benefit prosthetics users. Therefore, as an effort to transfer dexterous space robotics technology to prosthetics applications and to benefit from existing myoelectric technology, NASA is collaborating with the Limbs of Love Foundation, The Institute for Rehabilitation and Research, and Rice University in developing improved myoelectric control of multifinger hands and prostheses. In this paper, we will address the objectives and approaches of this collaborative effort and discuss the technical issues associated with myoelectric control of multifinger hands. We will also report our current progress and discuss plans for future work.

INTRODUCTION

Robotics is one of the critical technologies necessary for future space explorations. Future space robots will require highly dexterous robotic hands to perform a variety of tasks. A major technical challenge in making this possible is an improvement in the way these dexterous robotic hands are remotely controlled or teleoperated. The required robotic hand teleoperation interface must be intuitive (requiring less operator training) and nonfatiguing (enabling longer shifts). A current method of teleoperation uses an exoskeleton glove controller to detect finger motions. These glove controllers are worn by an operator to control robotic hands located at a remote site. Glove controllers are usually bulky and heavy and sometimes interfere with hand movements. Consequently, NASA Johnson Space Center (NASA/JSC) is investigating the feasibility of using myoelectric signals to control dexterous robotic hands.

While NASA is advancing dexterous robotic hand technology, the Limbs of Love Foundation, a foundation dedicated to providing prostheses to handicapped children, is actively searching for ways to improve the state-of-the-art in prostheses. In an effort to transfer advanced space technology to practical ground-based applications, NASA has teamed up with Limbs of Love and a group of medical and prosthetics specialists, prosthetics users, insurance industry representatives, and university researchers to identify research objectives in prosthetic hands [32]. As part of this effort, the Automation and Robotics Division (A&RD) at NASA/JSC has been actively working with Rice University to improve dexterous hand design and to develop a method for myoelectric control of multifinger hands.

This paper describes the collaborative research between NASA/JSC and Rice University in developing improved prostheses. First, the paper reviews previous work in dexterous robotic hands, prosthetics, and myoelectric controls, then it outlines the goals and objectives as well as the approaches we are taking in this joint effort. This paper also reports progress we have made in the areas of dexterous robotic hand development and myoelectric control. Our efforts in these areas have forced us to consider several difficult design issues which will be discussed. Finally, the paper concludes by stating what our future work and expected accomplishments will be.

PREVIOUS WORK

Over the past three decades, the myoelectric prostheses community reached a rough consensus that there are five types of grasp important in a person's daily activities: (1) three-jaw chuck or pincher grasp used to hold small objects; (2) lateral grasp, most often called a key grasp because it is used to hold a key while unlocking a door; (3) hook grasp, used to carry items such as books or a briefcase; (4) spherical grasp, where the thumb and fingers are wrapped around a spherical object; and (5) cylindrical grasp, where the thumb and fingers are wrapped around a cylindrical object [28] [40] [41]. Some consider the flattened hand (with thumb rotated completely out of opposition of the fingers) a sixth grasp, as it is essential in supporting flat objects such as trays.

Current commercial prostheses have a two-jaw pincher arrangement of fingers and thumb which gives some chuck and cylindrical grasping capability. More advanced prostheses, such as those described in references [12], [13], [27], and [34], have incorporated chuck and key grasps with spherical and cylindrical grasp options being provided by passive finger compliance. Weight, size, cost, and reliability of these advanced prostheses have been major reasons why they never became commercial products; however, recent advances in miniaturizing hardware and lowering power consumption and costs suggest that these problems may now be secondary to the control/user interface problem. In fact, the longest (over 15 years) multifunction prosthetic hand project, the Swedish hand, ended with this conclusion [2].

In parallel with the prosthetics research effort, the robotics community has been developing the theory of grasping and manipulation by multifingered hands over the past decades. (See, for example, the books by Mason and Salisbury [36] and Cutkosky [6]). Some multifingered robotic hands have been constructed. Hess and Li [16] provided an overview of several existing dexterous robotic hands for space applications, with the most dexterous of them being the Utah/MIT Dexterous Hand (UMDH) [23] and the Stanford/JPL Hand [36]. Based on their evaluations, Hess and Li have concluded that a six degree-of-freedom (DOF) robotic hand has sufficient articulation for grasping various shapes and providing some manipulation capability. Unfortunately, these early hands are too bulky and heavy to be feasible as a prosthetic device.

Although these complex robotic hands may not be feasible for prosthetics applications, they do serve as valuable tools to evaluate various grasping and manipulation strategies. These strategies usually involve complex algorithms and require sophisticated sensors now unavailable. One promising approach has been suggested by Speeter [39]. He uses a small set of basic grasping primitives, each of which is simple to program. This approach seems well-suited for application to teleoperation using probably a small set of myoelectric signals. Since each myoelectric input signal requires amplification, filtering, and processing, fewer inputs mean a less complex and less expensive user interface.

RESEARCH OBJECTIVES

Our review of previous work in dexterous robotic hand and myoelectric control has shown that to develop an improved prosthetic hand, progress must be made in (1) increasing the articulation of prostheses beyond just a single DOF, and (2) improving myoelectric sensing capability to recognize different muscle patterns and map them into various grasp primitives. To achieve progress in these two areas, we must accomplish the following set of objectives:

- Design a robotic hand with human compatible functions, weight, and size.
- Develop electronics and algorithms for primitive-based hand control.
- Develop a myoelectric pattern recognition technique for multifinger hand control.

The first objective will make a dexterous robotic hand feasible for limb-deficient persons. This may mean increasing the number of active fingers and reducing weight and power consumption. The second objective is aimed at providing some local automation so that the user can interface with the hand using primitive-level commands (such as chuck grasp, key grasp, etc.). To differentiate one primitive command from another, we must be able to recognize the signature, or myoelectric pattern, associated with that particular primitive. Previous attempts to develop more capable myoelectric prosthetic hands have fallen victim to an inadequate myoelectric user interface. This is why the third objective is so essential.

By achieving these three objectives, both NASA and the prosthetics community will benefit from the results of this effort. In space, electrical power is a precious commodity and weight is a major concern; by making the robotic hands more compact and energy efficient, we are also making them more suitable for space applications. The electronics and algorithms for primitive-based control will fit in well with a layered architecture that also supports artificial intelligence technologies. If the third objective is also achieved, it will represent a breakthrough in teleoperation technology since cumbersome exoskeleton devices will not be required to operate dexterous robotic hands.

APPROACHES

To achieve the three objectives, we are pursuing two parallel paths: JSC is focusing on developing advanced dexterous robotic hands, and Rice University is concentrating on developing improved myoelectric signal processing technology. In addition, we also solicit feedback from robotic hand experts, prosthetics users, and specialists to continually improve our design. These two technology development paths will eventually merge in a test bed environment where we can perform integrated evaluation of new prosthetics mechanisms and control.

DEXTEROUS ROBOTIC HANDS

We began dexterous robotic hand development by procuring and evaluating commercially available, state-of-the-art dexterous robotic hands while developing in-house expertise. By understanding the features of existing hands, we would not have to reinvent the technologies already developed by others. The results of our evaluation helped us to understand the trade-off between function (dexterity, sensing) and form (size, weight).

To establish a reference point for our performance evaluation, we first examined conventional parallel jaw grippers. Conventional parallel grippers are typically designed to execute a pinch grip. This type of grip depends heavily on contact friction rather than contact geometry for stability. Most grippers today have only a single DOF; therefore, they cannot perform manipulation or securely grasp objects of various shapes. Prosthetic hands today function essentially like parallel jaw grippers, except that prosthetic hands generally have a more human-like external appearance.

CTSD I Hand

For robotic applications, we took a minimalist approach in designing new hands. Instead of designing a highly complex robotic hand right away, we increased the complexity slowly, hoping to achieve the desired functions at a minimal cost. Our first attempt at robotic hand design resulted in the construction of the CTSD I Hand. (CTSD stands for Crew and Thermal Systems Division, the NASA/JSC organization responsible for developing the hand.)

The CTSD I Hand, shown in figure 1, has three fingers driven by a single DC motor. The three fingers are spaced 120 degrees apart, and they open and close simultaneously. Each finger contains three sections connected by joints. The sections are coupled by direct linkages; therefore, the push-pull motion created by the rod inside the proximal finger section will cause the other sections to move also. As the fingers begin to close, the distal finger section will bend around the object and trap the object within the grip of the hand for a secure grasp. The motions of the three fingers are also coupled by a cable-pulley system, so when any one finger is forced to stop, the other two will continue to close until all three fingers have stopped. Although this hand is a step beyond the simple parallel jaw gripper, it still has some drawbacks. The hand does not have enough independently controlled, articulated joints to allow alternate grasp arrangements, and it lacks the human look that is highly desired in a prosthetic hand.

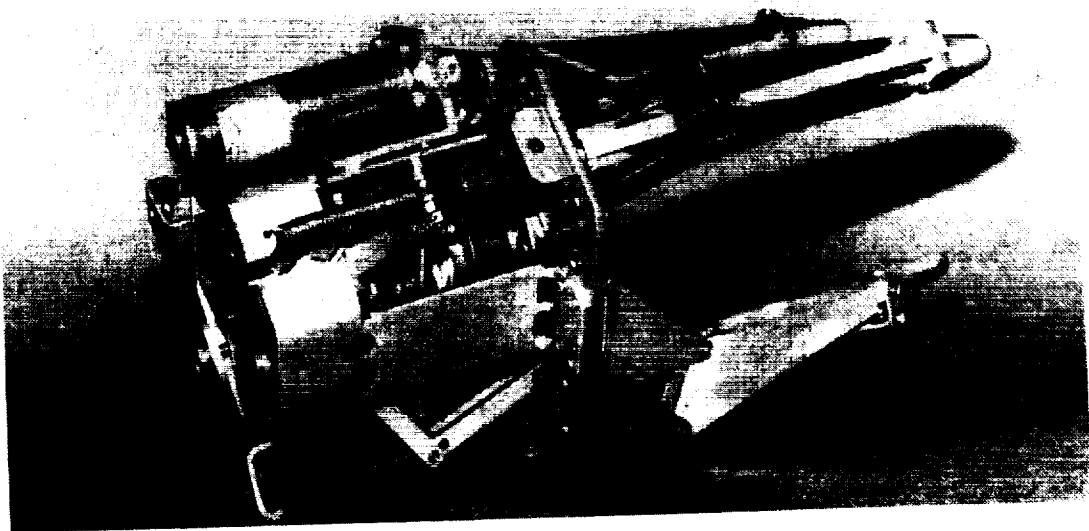


Figure 1. CTSD I Hand.

CTSD II Hand

To improve the dexterity of the CTSD I Hand, we redesigned the fingers so they are modularized, each capable of moving independently from other fingers. The redesigned hand, named the CTSD II Hand, also has three fingers. However, there are several important differences. The fingers of the CTSD II Hand, shown in figure 2, are arranged in a two-opposing-one configuration to provide parallel grasping surfaces. This finger configuration is able to adapt to different shapes of objects better than the CTSD I Hand configuration. The modular finger design also allows additional fingers to be added if necessary. Each finger is driven by a single DC motor contained within the finger module. We also introduced tactile sensors and strain gauges on each finger to provide sensory feedback [16]. Silicon pads cover the tactile sensors for protection and provide a compliant, friction surface for a more secure grasp. The maximum amount of force each finger can exert is controlled by current-limiting circuitry in the control electronics. The CTSD II Hand contains many functional improvements over the CTSD I Hand. However, for prosthetics applications, the CTSD II Hand lacks adequate dexterity and a pleasing appearance.

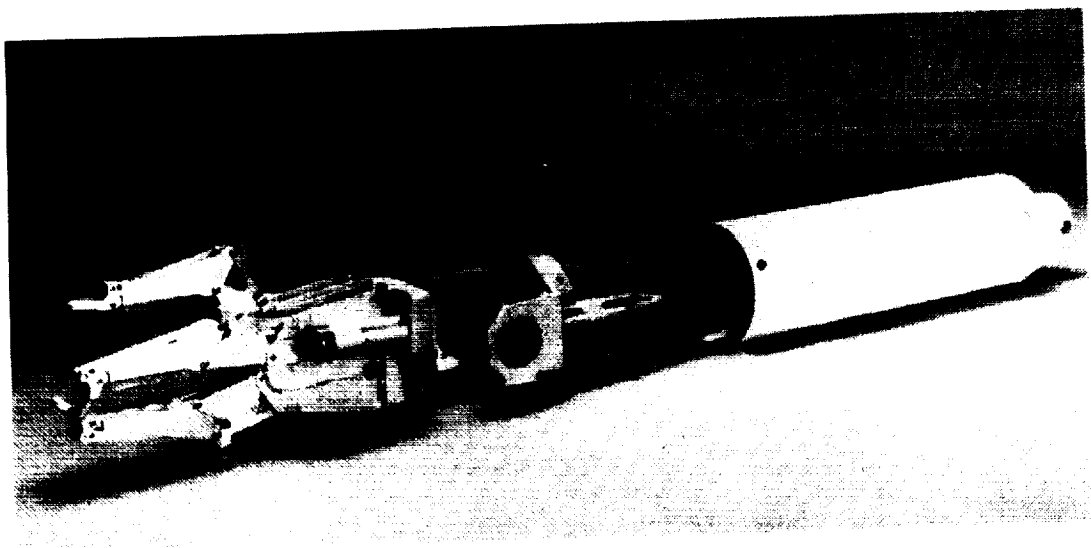


Figure 2. CTSD II Hand.

Utah/MIT Dexterous Hand

Our search for a robotic hand with human-like dexterity and appearance led us to evaluate the Utah/MIT Dexterous Hand (UMDH) [24]. The UMDH, shown in figure 3, is the most dexterous hand in the spectrum of hands available for our evaluation. It has 16 DOF arranged in an anthropomorphic configuration of three fingers and a thumb. The fingers and the thumb each have 4 DOF. Thirty-two pneumatic actuators operating at pressures up to 80 psi provide power to the hand. Tendons are used to transmit power from these pneumatic actuators to the joints through a system of pulleys and linkages called a "remotizer." Each joint is controlled by a pair of antagonistic tendons. Located inside each joint is a linear Hall Effect sensor that measures the joint angles. Hall Effect sensors are also located in the wrist to monitor the tendon tensions. A control box containing analog feedback control circuitry provides manual control of each joint with an interface for computer control that can be used in lieu of manual control [16].

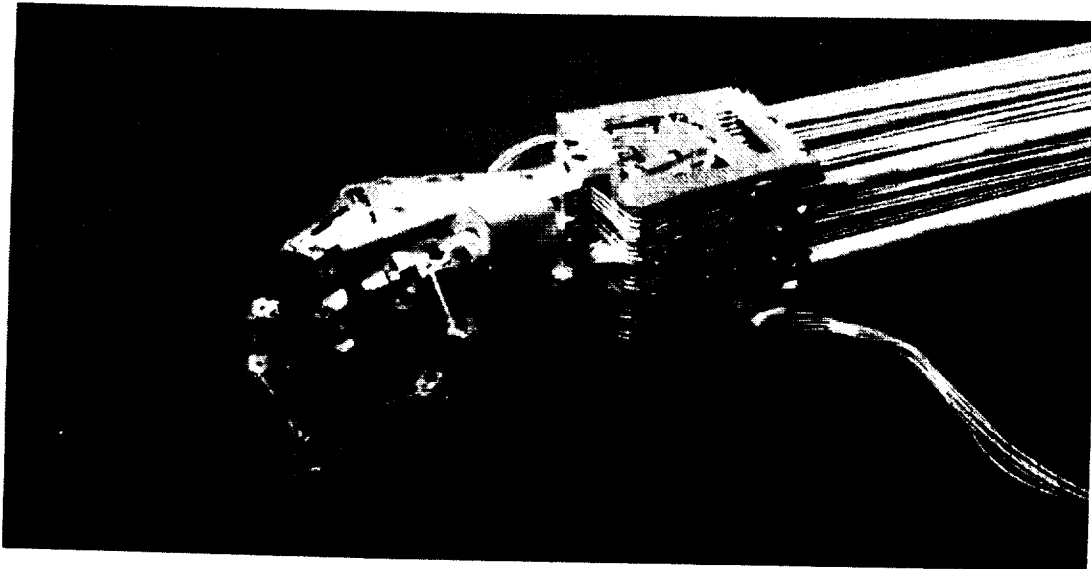


Figure 3. Utah/MIT Dexterous Hand.

It is obvious that the UMDH is not suitable for space robotics or for prosthetics applications. The pneumatic power system requires an air compressor too large to be portable, and the overall dimension of the hand system is too large to be mounted on a robot or a human user. However, the UMDH is a valuable test bed facility for us to evaluate and develop various control algorithms and grasp strategies for space and prosthetics applications. Later in this paper, we describe how the UMDH test bed is being used for prostheses development.

Stanford/JPL Hand

We also evaluated the Stanford/JPL Hand designed by Dr. J. Kenneth Salisbury of Massachusetts Institute of Technology (MIT). The hand has 9 DOF in a nonanthropomorphic finger arrangement and a large envelope of excursion. The hand has three fingers, each with three joints. The joints are driven by a set of steel cables that transmit mechanical power from 12 remotely located DC motors equipped with position encoders. Located behind the proximal joint of each finger are four strain gauges that measure the cable tensions. The tension signals may be translated into joint torque signals which are used in servo control. The fingertips are made of a highly compliant elastomer that provides the friction contact necessary for a secure grasp. Figure 4 shows the Stanford/JPL Hand and its remote motor package.



Figure 4. Stanford/JPL Hand.

Compared to the UMDH, the Stanford/JPL Hand is more compact, and its electrical power system is more compatible with space and prosthetics applications. While the size and weight of the Stanford/JPL Hand are acceptable for space robots, they are not acceptable for a prosthetic hand. Also, the Stanford/JPL Hand is not as anthropomorphic and visually pleasing as the UMDH.

Direct Link Prehensor

Our initial evaluation of the UMDH and the Stanford/JPL Hand showed us that a highly complex robotic hand will most likely require a large actuator package. This is unacceptable for both space and prosthetics applications; however, a smaller actuator package usually means less dexterity. Therefore, a compromise must be achieved between dexterity and packaging. Our search for an optimal solution that takes both packaging and dexterity into account brought us to the Direct Link Prehensor design.

The Direct Link Prehensor, as shown in figure 5, was originally developed by NASA Ames Research Center and Stanford University to function as a space suit end effector that fits over the hand like a glove. The prehensor has a total of 6 DOF in an anthropomorphic configuration. It has two fingers and a thumb, with the thumb opposing the two fingers at a fixed angle to provide grasping capability as well as some manipulation capability. The mechanical fingers are directly coupled to their human counterparts through a mechanical linkage system.



Figure 5. Direct Link Prehensor.

The prehensor has been flown on the NASA KC-135 aircraft to evaluate grasping in a weightless environment using a mechanical hand [25]. This evaluation showed the prehensor's finger arrangement to be a good compromise between packaging and dexterity. A robotic implementation of the prehensor would require only six motors, which is substantially less than the UMDH and Stanford/JPL Hand. Even with only 6 DOF, the prehensor is capable of grasping objects of various sizes and shapes. It is capable of chuck grasps, pinch grasps, power grasps, hook grasps, and key grasps. Although the thumb does not have abduction/adduction movement, it is mounted at a 45-degree angle to provide a motion with a horizontal component. Despite lacking some important movements, we were able to twist open a bottle cap, manipulate small flat plates, grasp balls and cylinders, and pick up luggage with the prehensor.

JH-3 Hand

After a fairly comprehensive evaluation of existing dexterous robotic hands, we selected the Direct Link Prehensor design as the baseline for an in-house developed robotic hand. After several design iterations (JH-1 through JH-2), we arrived at the JH-3 Hand. (JH stands for Jameson Hand, named after the designer Dr. John Jameson.) As shown in figure 6, the JH-3 Hand has an integrated hand-wrist-forearm package that approximates the combined size of a human hand, wrist, and forearm. Seven DC motors are packaged in the forearm: one motor per each DOF and one that controls the tendon tension. The wrist on the JH-3 Hand comes from a Remotec RM-10A robotic arm. Power is transmitted from the motors through a tendon-pulley system to each joint, much like the remotizer in the UMDH. This tendon-pulley system allows the hand to move freely with the wrist. The encoders on each motor and the strain gauges in the hand provide position and force feedback. Infrared proximity sensors were installed on the JH-3 Hand to provide autonomous adaptive grasping capability. The entire hand package contains current drivers for the motors as well as signal amplifiers for the sensors. Although the overall weight and package are still not quite acceptable (15 lbs), the JH-3 Hand does contain major improvements in packaging and sensing as compared to the UMDH and Stanford/JPL Hand.

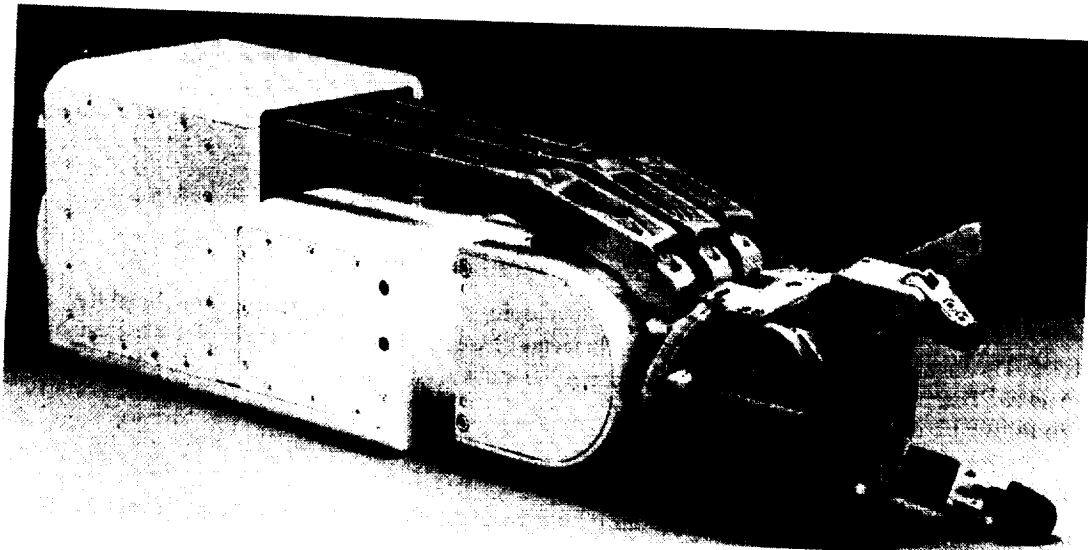


Figure 6. JH-3 Hand.

To evaluate the JH-3 Hand, we mounted the hand on the EVA Retriever, an in-house developed, highly autonomous, free-flying robot which operated on an air-bearing floor at NASA/JSC. From our evaluation, we arrived at two key conclusions related to prosthetics development. First, the mechanism for "remotizing" the actuators tends to add weight, bulk, and complexity to the overall system. Instead, local actuation requiring a minimum number of power transmission components is desired. There is a design trade-off between remote actuation and local actuation. Remote actuation adds to the overall weight of the system, but allows a more desirable mass distribution for moment reduction. On the other hand, local actuation tends to concentrate mass near the hand and amplifies moments about the elbow and

shoulder. Second, the integrated hand-wrist-forearm design does not permit a simple integration of the JH-3 Hand with commercial robotic arms. Since most commercial robotic arms already come with a forearm fully integrated, installing a JH-3 Hand on these robotic arms requires redesign.

JH-4 Hand

Incorporating the lessons learned from the JH-3 Hand evaluation, we developed the most recent hand design: the JH-4 Hand. This hand, shown in figure 7, contains two fingers and a thumb, each driven by two motors located right behind the proximal joint. Instead of remotizing the motors and transferring mechanical power through tendons and pulleys like the JH-3 Hand, motors in the JH-4 Hand drive the finger joints directly with a minimum number of gears. In making the fingers truly modular, we also packaged the drive electronics (e.g. current amplifiers, motion controllers) into each finger. An 80C196 microcontroller provides each finger with some local intelligence and serves as a high-level command interface.

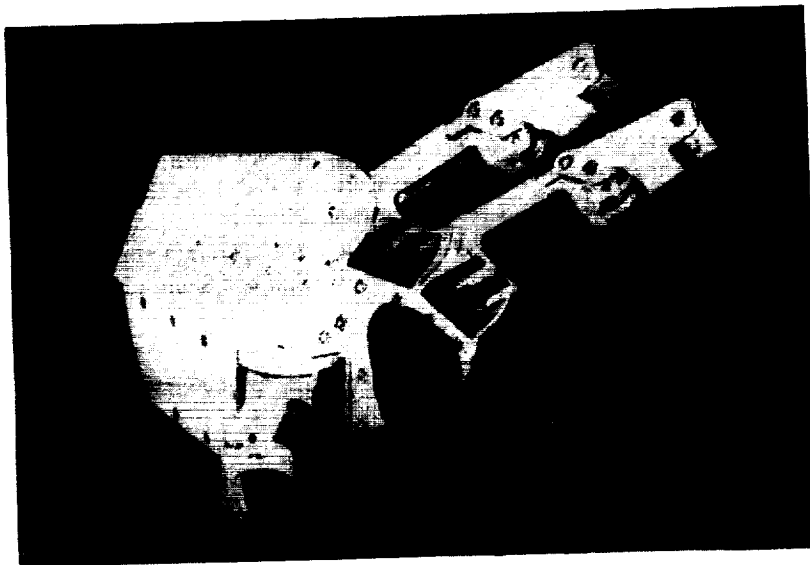


Figure 7. JH-4 Hand.

The JH-4 Hand represents our latest effort in developing a modular dexterous robotic hand that satisfies the stated objectives. The hand is near human-equivalent in terms of weight and size, and provides a reasonable degree of dexterity with its two fingers and a thumb. The microcontroller embedded in each finger provides a high-level command interface for primitive-based hand control.

Rice Prosthetic Hand Prototypes

In parallel with these NASA/JSC dexterous robotic hand developments and with technical consultation from NASA A&RD experts, Rice University engineers have begun developing prototype anthropomorphic prosthetic hands. Two hands and one wrist unit have been built. Each hand has a thumb and four fingers with independent thumb and finger motion. One hand has a single-axis thumb which allows the thumb to swing through a full range of opposition to the finger tips and independent finger control on the index, middle, and ring fingers; the little finger is coupled to the ring finger. The other hand has a two-axis thumb which can abduct, oppose, and flex at its base; an independent index finger; and the remaining fingers coupled to complete grasps. Both hands can perform key, chuck, cylindrical, spherical, and hook grasps as well as completely flatten. The wrist unit is capable of flexing and roll. We are now beginning our second design iteration to simplify and strengthen the mechanisms to make them more reliable and easier to manufacture.

MYOELECTRIC CONTROL

To achieve our third objective, we are investigating the feasibility of using myoelectric signals to control the robotic hands. Our first efforts are exploring the remote control or teleoperation scenario. Myoelectric teleoperation of dexterous robotic hands will require no mechanical parts, and may greatly reduce the bulk and weight now found in dexterous robotic hand control devices; however, this teleoperation scenario requires advances in the state of myoelectric control art. Improvement in myoelectric teleoperation of multifingered hands will also benefit the prosthetics users. For example, the level of myoelectric control of a dexterous hand achieved by an intact teleoperator will establish an upper performance bound for the amputee. Furthermore, some of the myoelectric signal processing techniques developed for teleoperation will transfer into prosthesis control.

Research in using myoelectric signals (also called electromyographic or EMG signals) to control prostheses dates from the late 1940's [35]. By the early 1970's, researchers were treating the myoelectric signals as an amplitude-modulated signal whose amplitude was roughly proportional to the force developed in the muscle generating the myoelectric signal. The consensus was that most of the information in a myoelectric signal was in the amplitude [18]. By the late 1970's, the model had matured to treating the myoelectric signal as amplitude-modulated Gaussian noise whose variance was proportional to the force developed by the muscle [31] [37].

Today's commercial myoprocessors used in prosthesis control are based on only one dimension of the myoelectric signal, the force level, and in a few cases, its rate of change. Researchers have successfully refined force estimation from the myoelectric signal [3] [5] [19] [24] [30] [31] [37]. Parker's work forms the basis of control multiple functions using different force levels on a signal channel [37]. Hogan's work was particularly significant in eliminating low frequency noise from the force estimates due to the spatio-temporal sampling artifact inevitable with skin surface electrodes [18] [19]. Jacobsen [24] refined use of the rate of change of force in elbow control of the Utah arm. A version of the Swedish Hand used rate of change of force to switch control functions [13]. Rice University researchers have investigated these force estimation results in operating a proportionally controlled grasp force with a three-fingered robotic hand.

These force-estimation techniques require a separable muscle contraction for each function commanded, making simultaneous control of two or more joints very difficult. A number of researchers, beginning with Wirta and Taylor [42], examined linear combinations of myoelectric force estimates from multiple channels to select different functions. The Swedish Hand developers applied these methods to selecting wrist and grasp [1] [2] [15]. The Japanese research team applied the technique to wrist control in the Waseda Hand 3 [27]. Jacobsen [22] and Jerard [26] formalized the mathematics for this approach and applied it to upper limb above-elbow prostheses. These force-estimating approaches require at least one electrode pair and signal processing channel for each muscle used, up to a dozen in some above-elbow experiments. Furthermore, force-estimating myoprocessors can be used only on superficial muscles [19], while most motions involve both superficial and deep muscles. In fact, any deep muscle activity reaching a force-estimating myoprocessor is mistakenly interpreted as superficial muscle activity. Therefore, it appears to us that to obtain multifunction sensitivity that is intuitively easy to use, all information in the myoelectric signal must be exploited, rather than just the force estimate. In addition to using superficial muscles (to which force-estimation techniques are limited), the deep muscles must also be used in myoelectric control systems.

Some researchers have considered shape and spectral characteristics of the myoelectric signal in addition to force estimation. Recent findings suggest that there is considerable information in the myoelectric spectra, if we can understand its coding. Examples include:

- Small muscles generally have fewer fibers per single motor unit (SMU) and therefore have power spectra containing more high-frequency activity than larger muscles with larger SMUs [33].
- Tissue (including other muscles) between the active muscle and the measuring electrode acts as a low pass filter to myoelectric signals, thus excessive low-frequency power densities may indicate cross talk from adjacent muscles [33].
- Action potential conduction velocity decreases with fatigue, causing gradual shifts in power from higher to lower frequencies during sustained forceful contractions [33].
- SMU recruitment order is stable for a given task [7]. Short-time spectra of myoelectric signals associated with a given rapid movement does not vary as much as previously thought [14].

The full spectrum of the myoelectric signal has been examined using techniques involving statistical pattern and spectral analyses [8] [9] [10] [11] [29] [38]. Evidence of movements having distinct spectral signatures has been reported by Lindstrom and Magnusson [33], DeLuca [7], and Hannaford and Lehman [14]. The spectral signature of the initial muscular recruiting phase of arm motions to select up to six functions of an upper limb prosthesis from a single myoelectric signal has also been exploited recently by Hudgins [20].

Hudgins' [20] use of spectra-related parameters, such as zero-crossing and slope changes, and the use of Short Time Fourier Transforms by Hannaford led us to focus on the time-varying spectrum of the myoelectric signal in our research. We have been studying the correlation between the myoelectric spectrum in the initial recruiting phase of a motion with the type of motion. We are following the lead of Saridis [38], Doerschuk [8], Kelley [29], and Hudgins [20] in using the traditional single-muscle signals. However, our work differs from previous work of other researchers in that we are using the actual frequency spectrum to discriminate different grasping motions. Also, previous work has focused on arm, not hand, motions and on parameters derived from the spectrum rather than the actual spectrum.

Myoelectric Experimental Setup

To evaluate various myoelectric control techniques, we developed a unique myoelectric data collection system which enables us to capture up to eight myoelectric data streams while simultaneously recording the motion of the subject's hand. Previous myoelectric researchers have had only limited, if any, capability to measure motion while measuring myoelectric signals.

Figure 8 is a block diagram of the data capture system. We use the Dexterous Hand Master (DHM), an exoskeleton glove manufactured by EXOS, Inc., of Cambridge, Massachusetts, to measure the subject's joint angles. The DHM glove, also used by NASA/JSC as a master to teleoperate dexterous robotic hands, measures parameters related to joint angles for four joints on the thumb and each of three fingers (index, middle, and ring) [43].

We use the Grass Instruments (Quincy, MA) Model 12 amplifier to measure myoelectric signals. It consists of a differential amplifier, a high-pass filter (with roll-off frequency adjustable from 0.01 to 300 Hz) to block DC and motion artifact, a low-pass filter (adjustable from 30 to 20,000 Hz) to limit aliasing, an adjustable gain amplifier stage, and an isolation to protect the subject from the electric shock hazards of power supply and computer equipment. This sequence amplifies the differential myoelectric signal from skin-surface electrodes (around 1 millivolt in amplitude) to several volts. Differential input reduces the 60 Hz interference (typically much larger than the myoelectric signal) from lights and equipment.

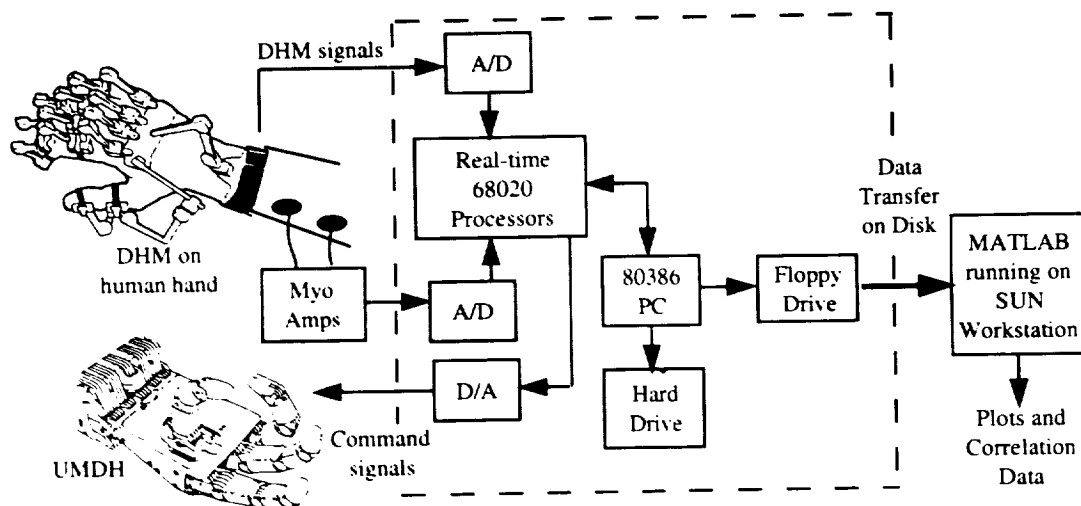


Figure 8. Myoelectric data capture system.

Both DHM and myoelectric amplifiers are connected through Burr-Brown MPV950S analog-to-digital converter (ADC) boards to 68020-based Ironics IV3204 and IV3201 microcomputers. These capture up to 32 channels of data at 1000 samples per second per channel. An 80386 Radix PC transfers data to MATLAB format disk files. The ADC and Ironics and Radix computers are on a VME bus and do double duty as the control computer for the UMDH, with which we plan to demonstrate myoelectric teleoperation. We used Math Work's (Natick, MA) MATLAB software, Version 4.1, for off-line data analysis and plotting.

Based on our early findings in the myoelectric spectra, Rice University also developed a real-time myoelectric control implementation test bed consisting of miniaturized myoelectric amplifiers with fixed 20 to 500 Hz bandwidth and a personal computer incorporating an Elf-31(TM) board and signal processing development software by Atlanta Signal Processors. The Elf-31 includes high-speed analog-to-digital conversion and digital signal processor (TMS320C31). With this system, we can capture multiple myoelectric signals, compute their spectra on the Elf-31, and process the spectra into a grasp selection using the personal computer, all in real time. NWorks (TM) software by Neural Ware, Inc. (Pittsburgh, PA) facilitates development of neural networks for the classification of the spectra, if the user chooses a neural network approach.

Preliminary Myoelectric Results

Our first use of these systems investigated direct use of the myoelectric spectrum to differentiate the key and chuck grasps. The key and chuck grasps differ in thumb position relative to the fingers. The thumb opposes the side of the index finger in the key grasp, while it opposes the tips of the index and ring fingers in the chuck grasp. Anatomy suggests that differentiating between these grasps requires measuring intrinsic thumb muscle activity in the hand and extrinsic finger and thumb activity in the forearm. We restricted measurements to the forearm, however, to keep the teleoperator's hand free of movement-encumbering hardware and increase our work's applicability to the prosthetics community.

We used the NASA/JSC data collection system (mentioned earlier) to capture myoelectric data during these two grasps. The human subject did a series of key and chuck grasps with the lateral side of the forearm resting on a horizontal surface or hanging vertically. We made no attempt to control the starting position (a relaxed posture) or grasp precisely. The subject was the judge of consistency in these positions. We later used DHM finger trajectory data to check for consistency and correctness of the grasp and to locate the initiation phase of the corresponding myoelectric signals.

We are now testing various myoelectric signal processing schemes on these data streams. One was an adaptation of the approach by Hudgins et al., [20] at the University of New Brunswick, where they computed mean absolute value, mean absolute value slope, zero crossings, and waveform length on biceps and triceps in 40 ms windows in the first 240 ms of arm motions, such as elbow flexing and humeral rotation. A multilayer Perceptron neural network used these features to classify the arm motions with 70 to 98% accuracy, depending on the human subject. Our initial implementation of this scheme yielded a maximum of 80% correct accuracy for our grasping test set. We believe that the decreased accuracy may be due to (1) increased difficulties in detecting the grasp start (since the myoelectric signal amplitude is smaller) and (2) differences in the way muscles used in fine motion control (such as grasping) and coarse motion (such as arm motion) are recruited. Inaccuracies due to the latter may be reduced by reoptimizing window size and characteristics. More algorithm experimentation and trials on more human subjects are needed to confirm this, however.

We have also tested several schemes which use the myoelectric signal's magnitude spectra directly. The most successful of these have used the upper portion of the myoelectric spectrum, the 75 to 250 Hz range. Muscle fiber length, diameter, and action potential conduction velocity as well as distance to the electrode dominate this portion of the myoelectric spectrum. A multilayer Perceptron receiving inputs of the 75 to 250 Hz spectrum in six 40 ms windows (as in the UNB scheme) from the distal channel in figure 9 classified the test set signatures 93% correctly.

We have also experimented with multiple channel configurations. Figure 9 shows our dual channel electrode configuration; the distal electrode pair (D1 and D2) measure extrinsic thumb muscle activity while the proximal pair (P1 and P2) measure finger flexion and extension activity. Computing 6 values of the 75 to 250 Hz spectrum in larger (240 ms) windows on both channels during the motion's initiation phase yields a set of 12 features that a multilayer Perceptron can classify 86 to 91% correctly, depending on the human subject. We implemented this approach in real time on our Real-Time Myoelectric Control Implementation Testbed and used it to teleoperate the Rice-developed prosthetic hands described previously.

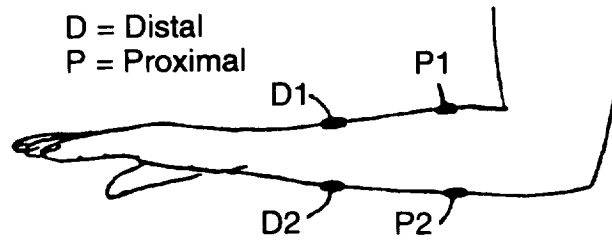


Figure 9. Dual channel electrode configuration.

While still below our grasp discrimination goal of 100%, these early results refute the long-held assumption that myoelectric signals from the forearm are inadequate for differentiating thumb motions.

We have begun experimenting with the 3 to 75 Hz portion of the myoelectric spectrum, which is dominated by muscle recruitment dynamics. Theoretically, task-specific SMU recruitment should show up in this portion of the spectrum, and this may be the key to increasing grasp selection accuracy above 93%. To date, we have not been successful with direct use of the magnitude spectrum in this region, in spite of the implied usefulness in references [7], [14], and [20] and the results of the adaptation of the UNB scheme, which implicitly used the entire spectrum. The UNB scheme has phase information embedded however. Since we expect the recruitment portion of the spectrum to be much more time-varying, it will be especially sensitive to window characteristics. We are continuing experiments with varying window size, overlap, and type.

CONCLUSIONS AND FUTURE WORK

This ongoing joint research effort between NASA/JSC and Rice University is entering its third year. In the past two years, we have made significant progress in accomplishing the three stated objectives. We have evaluated several commercially available dexterous hand designs and gone through several iterations of our own in-house designs. We made progress in reducing weight and packaging of dexterous robotic hands while maintaining an acceptable level of dexterity, and realized the current design in the JH-4 Hand. We have begun development of prosthetic hands that incorporate lessons learned from robotic hand design and control.

Meanwhile, we also made significant progress in understanding myoelectric control theory. We have developed a unique myoelectric data collection system featuring recording of joint motion, and developed a test bed for evaluating various signal processing techniques. Initial results of over 90% correct grasp discrimination suggest that myoelectric commanding of grasp primitives is feasible. Eventually, we plan to evaluate the feasibility of myoelectrically controlling individual fingertips to augment grasp primitives.

If myoelectric control of dexterous robotic hands can be made both intuitive to operate and repeatable, a myriad of opportunities in both space robotics and prostheses development will open up.

REFERENCES

- [1] Almstrom, C., *Myoelectric Control of Multifunctional Hand Prostheses - Contributions to the Pattern Recognition Approach, to Signal Acquisition, and to Clinical Evaluation*, Technical Report No. 79, School of Electrical Eng., Chalmers University of Technology, Goteborg, Sweden, 1977.
- [2] Almstrom, C., et al., "Experience with Swedish Multifunctional Prosthetic Hands Controlled by Pattern Recognition of Multiple Myoelectric Signals," *Int. Orthopaedics*, Springer-Verlag, Vol. 5, pp 15-21.
- [3] Bottomley, A., "Progress with the British Myoelectric Hand," *Proceedings of the Int. Sym. on External Control of Human Extremities*, Dubrovnik, Yugoslavia.
- [4] Brand, P., *Clinical Mechanics of the Hand*, The C. V. Mosby Co., St. Louis, 1985, pp 194-195.

- [5] Childress, D., "An Approach to Powered Grasp," *Proc. of the 4th Int. Sym. on External Control of Human Extremities*, Dubrovnik, Yugoslavia, Aug. 28 - Sep. 2, 1972.
- [6] Cutkosky, M., *Robotic Grasping and Fine Manipulation*, Boston, Kluwer, 1985.
- [7] DeLuca, C., "Physiology and Mathematics of Myoelectric Signals," *IEEE Trans. on Biomedical Eng.*, BME-26, No. 6, June 1979, pp 313-325.
- [8] Doerschuk, P., et al., "Upper Extremity Limb Function Discrimination Using EMG Signal Analysis," *IEEE Trans. on Biomedical Engineering*, BME-30, No. 1, Jan. 1983, pp 18-29.
- [9] Graupe, D., "Multifunctional Prosthesis and Orthosis Control via Microcomputer Identification of Temporal Pattern Differences in Signal Site Myoelectric Signals," *Journal of Biomedical Eng.*, Vol. 4, 1982, pp 17-22.
- [10] Graupe, D., et al., "A Microprocessor System for Multifunctional Control of Upper-Limb Prostheses via Myoelectric Signal Identification," *IEEE Trans. on Auto. Control*, AC-23, No. 4, Aug. 1978, pp 538-544.
- [11] Graupe, D. and Cline, W., "Functional Separation of EMG Signals via ARMA Identification Methods for Prosthesis Control Purposes," *IEEE Trans. of Systems, Man, and Cybernetics*, SMC-5, No. 2, Mar. 1975, pp 252-259.
- [12] Hagg, G. and Spets, K., "SVEN - Project I - Electrically Controlled Hand Prosthesis - Final Report," *Res. Inst. of the Swedish National Defence*, Dept 2, S-104 50 Stockholm 80 FOA 2 report A 2575-H5.
- [13] Hagg, G. and Oberg, K., *Adaptive EMG-Controlled Hand Prosthesis for Wrist Disarticulated Patients*, Ean-Holmgren Orthopaedic Co., Bergsbrunnagatan 1, S-753 23 Uppsala, Sweden, pp 441-449.
- [14] Hannaford, B. and Lehman, S., "Short-Time Fourier Analysis of the Electromyogram: Fast Movements and Constant Contraction," *IEEE Trans. on Biomedical Eng.*, BME-33, No. 12, Dec. 1986, pp 1173-1181.
- [15] Herberts, H., et al., "Hand Prosthesis Control Via Myoelectric Patterns," *Acta Orthopaedia Scandinavia*, Vol. 44, 1973, pp 389-409.
- [16] Hess, C. and Li, L. C., "Smart Hands for the EVA Retriever," *Proceedings of Third Annual Workshop on Space Operations Automation and Robotics (SOAR '89)*, pp 441-446, 1989.
- [17] Hlawatsch, F. and Boudreaux-Bartels, G., "Linear and Quadratic Time-Frequency Signal Representations," *IEEE Signal Proc. Mag.*, Vol. 9, No. 4, pp 21-67, 1992.
- [18] Hogan, N., "A Review of the Methods of Processing EMG for Use as a Proportional Control Signal," *Biomedical Engineering*, Mar. 1976, pp 81-86.
- [19] Hogan, N. and Mann, R., "Myoelectric Signal Processing: Optimal Estimation Applied to Electromyography," *IEEE Trans. on Biomedical Eng.*, BME-27, No. 7, July 1980, pp 382-410.
- [20] Hudgins B., et al., "A New Strategy for Multifunction Myoelectric Control," *IEEE Trans. on Biomedical Engineering*, No. 40, pp 82-94, 1993.
- [21] Ichie, M., et al. "EMG Analysis of the Thumb and Its Application to FNS," *IEEE 8th Annual Conf. of the Eng. in Med. & Biol. Soc.*, Fort Worth, TX, Nov. 7-10, 1986, pp 60-64.
- [22] Jacobsen, S., *Control Systems for Artificial Arms*, Ph.D. Dissertation, Dept. of Mechanical Eng., Massachusetts Institute of Technology, Jan. 1973.
- [23] Jacobsen, S., et al., "Design of the Utah/MIT Dexterous Hand," *IEEE International Conference on Robotics and Automation*, San Francisco, CA, Apr. 1986.
- [24] Jacobsen, S., et al., "Development of the Utah Artificial Arm," *IEEE Trans. on Biomedical Eng.*, BME-29, No. 4, Apr. 1982, pp 249-269.
- [25] Jameson, J., *Report on the Zero-G Grasp Experiment with the Direct Link Prehensor*, NASA/JSC, Mar. 15, 1988.
- [26] Jerard, R., *Application of a Unified Theory for Simultaneous Multiple Axis Artificial Arm Control*, Ph.D. Dissertation, Dept. of Mechanical and Industrial Eng., University of Utah, Dec. 1976.

- [27] Kato, I. and Okazaki, K., "Electro-Pneumatically Controlled Artificial Hand," *Bulletin of Science and Engineering Research Laboratory*, Waseda University, No. 44, 1969, pp 25-34.
- [28] Kay, H. and Rakic, M., "Specifications for Electromechanical Hands," *Proc. of the 4th Int. Sym. on External Control of Human Extremities*, Dubrovnik, Yugoslavia, Aug. 2 to Sep. 2, 1972.
- [29] Kelly, M., et al., "The Application of Neural Networks to Myoelectric Signal Analysis: A Preliminary Study," *IEEE Trans. on Biomedical Engineering*, BME-37, No. 3, Mar. 1990, pp 221-230.
- [30] Kreifeldt, J., "Signal Versus Noise Characteristics of Filtered EMG Used as a Control Source," *IEEE Trans. on Biomedical Eng.*, BME-18, No. 1, Jan. 1971, pp 16-22.
- [31] Kreifeldt, J. and Yao, S., "A Signal-to-Noise Investigation of Non-Linear Electromyographic Processors," *IEEE Trans. on Biomedical Engineering*, BME-21, No. 4, July 1974, pp 298-308.
- [32] Limbs of Love, "An American Initiative: The Next Generation of Myoelectric Prostheses," Workshop Report, NASA/JSC, Research Triangle Institute, 1991.
- [33] Lindstrom, L. and Magnusson, R., "Interpretation of Myoelectric Power Spectra: A Model and Its Applications," *Proceedings of the IEEE*, Vol. 65, No. 5, May 1977, pp 653-662.
- [34] Lozach, Y., et al., "On The Evaluation of a Multifunctional Prosthesis," *Proc. of the 7th World Congress of the Int. Soc. of Prosthetics and Orthotics*, Chicago, IL, June 28 - July 3, 1992, p 185.
- [35] Mann, R., "Cybernetic Limb Prosthesis: The ALZA Distinguished Lecture," *Annals of Biomedical Engineering*, Vol. 9, 1981, pp 1-43.
- [36] Mason, M. and Salisbury, J., *Robot Hands and the Mechanics of Manipulation*, Cambridge, MA, MIT Press, 1985.
- [37] Parker, P., et al., "Signal Processing for the Multistate Myoelectric Channel," *Proceedings of the IEEE*, Vol. 65, No. 5, May 1977, pp 662-647.
- [38] Saridis, G. and Gootee, T., "EMG Pattern Analysis and Classification for a Prosthetic Arm," *IEEE Trans. on Biomedical Engineering*, BME-29, No. 6, June 1982, pp 403-412.
- [39] Speeter, T., "Control of the Utah/MIT Hand: Hardware and Software Hierarchy," *Journal of Robotic Systems*, Vol. 7, No. 5, May 1990, pp 759-790.
- [40] Spinner, M., *Kaplan's Functional and Surgical Anatomy of the Hand*, Philadelphia, PA, JB Lippincott Co., 3rd Ed., pp 9-10.
- [41] Todd, R. and Nightingale, J., "Adaptive Prehension Control for a Prosthetic Hand," *Proceedings of the 3rd Int. Symposium on External Control of Human Extremities*, Dubrovnik, Yugoslavia, Aug. 25-30, 1969, pp 171-183.
- [42] Wirta, R. and Taylor, D., "Development of a Multiple-Axis Myoelectrically Controlled Prosthetic Arm," *Proc. of the 3rd Int. Sym. on External Control of Human Extremities*, Dubrovnik, Yugoslavia, Aug. 25-30, 1969.
- [43] Wright, A. and Stanisic, M., "Kinematic Mapping Between the EXOS Handmaster Exoskeleton and the Utah/MIT Dexterous Hand," *IEEE International Conference on Systems Engineering*, Pittsburgh, PA, Aug. 9-11, 1990, pp 101-104.

NAVY OMNI-DIRECTIONAL VEHICLE (ODV) DEVELOPMENT PROGRAM 2514
p. 10

Hillery McGowen
Coastal Systems Station, Dahlgren Division
Naval Surface Warfare Center
Panama City, FL 32407

ABSTRACT

The omni-directional vehicle (ODV) development program sponsored by the Office of Naval Research at the Coastal Systems Station has investigated the application of ODV technology for use in the Navy shipboard environment. ODV technology as originally received by the Navy in the form of the Cadillac-Gage Side Mover Vehicle was applicable to the shipboard environment with the potential to overcome conditions of reduced traction, ship motion, decks heeled at high angles, obstacles, and confined spaces. Under the Navy program, ODV technology was investigated and a series of experimental vehicles were built and successfully tested under extremely demanding conditions. The ODV drive system has been found to be applicable to autonomous, remotely, or manually operated vehicles. Potential commercial applications include multi-directional forklift trucks, automatic guided vehicles employed in manufacturing environments, and remotely controlled platforms used in nuclear facilities or for hazardous waste clean up tasks.

NAVY ODV DEVELOPMENT

The Navy ODV development program has investigated the application of ODV technology for use in the shipboard environment for cargo and ordnance handling. ODV technology as received by the Navy in the form of the Cadillac-Gage vehicle had the potential to overcome the limitations of existing vehicles under conditions of reduced traction, ship motion, decks heeled at high angles, obstacles, and restricted spaces. Development concerns focused on omni wheel complexity, footprint pressure, and traction; vehicle control, reliability, cost, maintainability, and autonomous and teleoperated vehicle control. Under the Navy program, a series of experimental vehicles were built and tested.

The omni wheel was originally patented by a Swedish inventor Bengt Ilon in 1973. The omni wheel and its operating principle is shown in Figures 1 and 2. The ODV is a four wheel drive system where each of the non-steerable wheels has its own drive motor. The omni-directional wheel allows the vehicle to travel in any direction, rotate about its axis, or to do both simultaneously (see Figure 3). Omni wheels are not steered as the plane of rotation is fixed in reference to the chassis. Mounted at a 45 degree angle to the wheel plane of rotation are a series of passive elliptical rollers. When a wheel is rotated, the resulting motion tends to move the wheel on the ground at a 45 degree angle to its plane of

rotation. By adding the individual motion created by each wheel, the vehicle can move in any desired direction. Vehicle speed and direction is controlled by a three axis joystick. Responding to the joystick, microprocessor-based algorithms control the rotation of each wheel to achieve the desired vehicle motion. The true three degree of freedom movement provided by the omni wheel, the capacity to operate under reduced traction conditions, and over substantial obstacles makes the ODV an ideal platform for robotic and teleoperated vehicles requiring high level control systems.

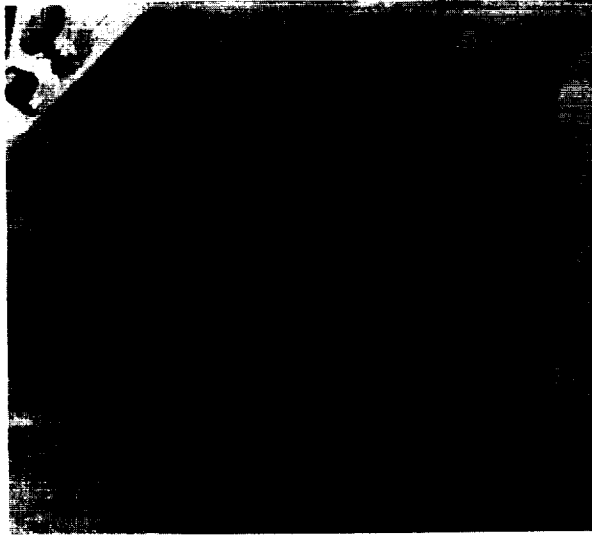


FIGURE 1. OMNI-DIRECTIONAL WHEEL

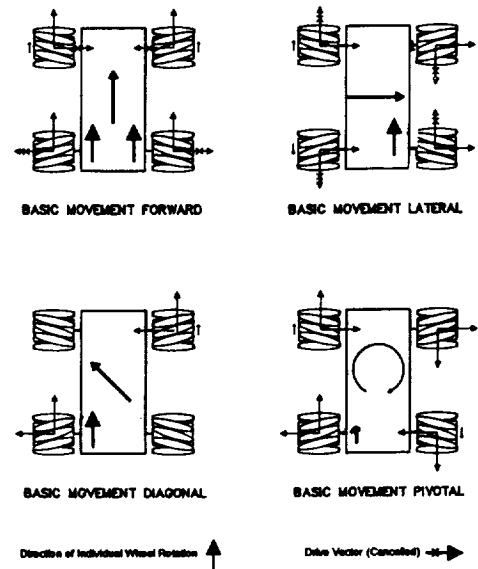


FIGURE 2. OMNI-DIRECTIONAL WHEEL OPERATING PRINCIPLE

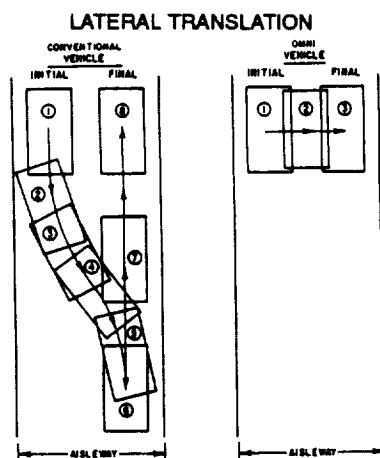
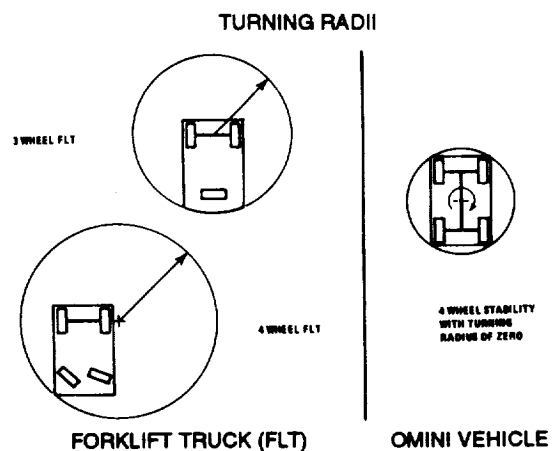


FIGURE 3. OMNI-DIRECTIONAL WHEEL MANEUVERABILITY VERSUS CONVENTIONAL VEHICLES



ODV AND ALL-WHEEL-STEERED VEHICLE COMPARISONS

Other high maneuverability drive systems including existing Navy multi-directional forklift trucks were compared with the ODV. This investigation found that the All-Wheel-Steered (AWS) vehicle was the only drive system capable of producing maneuverability approaching that of the ODV, and capable of being used in the shipboard environment. Two AWS sideloading forklift trucks have been used onboard Navy ships. It was established that the existing Navy AWS multi-directional vehicles had a number of limitations not shared by the ODV including:

- a. Limited maneuvering capability
- b. Mechanical complexity
- c. Difficult to operate and repair
- d. Scuffing damage to tires and deck non-skid surfaces

Complexity. The omni wheel is more complex than a conventional wheel. However, when conventional wheels are coupled to a suspension, drive, and steering system to produce a AWS vehicle, the vehicle is more complex than the ODV. Thus when addressing complexity and reliability, both wheel and vehicle complexity must be considered. With the exception of the wheel, the design and fabrication of an ODV is straightforward with most of the components available off-the-shelf. Mechanically, the ODV is a uncomplicated system: four identical drive units; the omni wheels; a simple suspension system; and a battery or diesel engine power source. Figure 4 illustrates the relative simplicity of the ODV mechanical and electrical design.

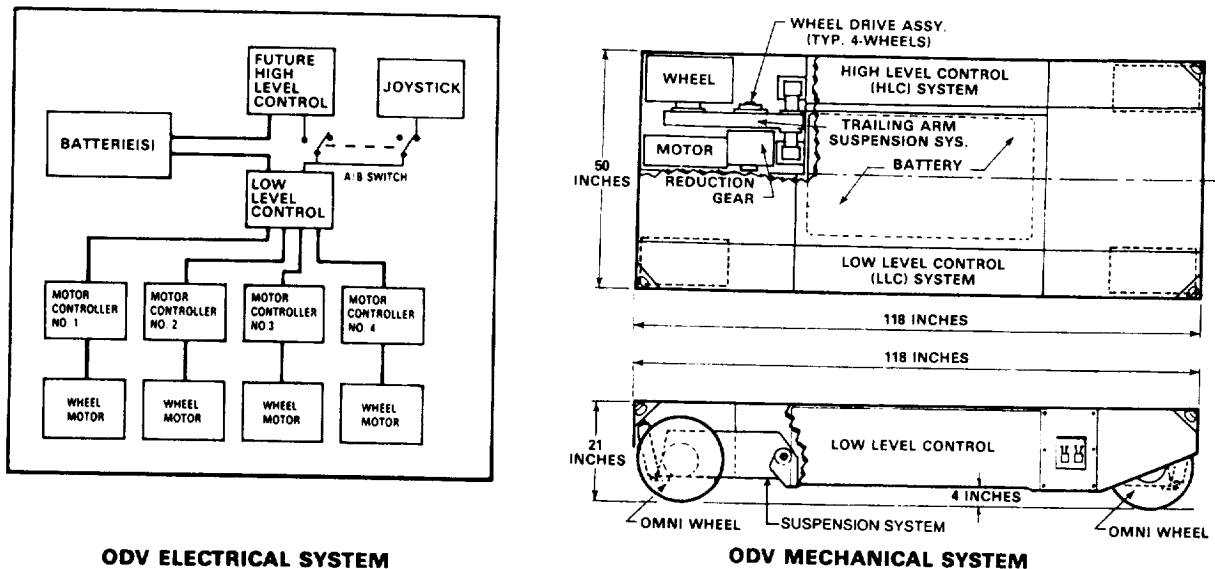


FIGURE 4. ODV MECHANICAL AND ELECTRICAL SYSTEMS

Maintenance. Maintenance for ODV and AWS vehicles must be considered on a system basis rather than on the wheel alone. The 19-year old Cadillac-Gage vehicle has never suffered a failure associated with the wheel even after having been run under adverse conditions including: sand, water, and mud. Three factors concerning wheel wear and failure should be noted. First, the omni wheel is essentially non-scuffing. Second, the time that the individual roller contacts the ground is only a portion of a revolution of the wheel. Third, a vehicle can be operated with a drive unit inoperative and the wheel free to rotate or even with a wheel removed, if the vehicle is suitably balanced.

The ODV drive system as seen in Figure 4 consists of four identical drive units; omni wheels, simple suspension system, a battery, and compact, easy to troubleshoot control electronics. Taken together these components equate to a low maintenance system. The omni wheel is more complex than a conventional wheel; however, the rest of the ODV is robust and straightforward. Conversely, steering, drive, and control mechanisms of the AWS vehicle are complex and present a continuing maintenance burden.

NAVY ODV PLATFORMS

A family of conceptual ordnance and cargo handling vehicles based on the omni drive system was developed by the Coastal Systems Station. An 8,000-pound capacity sideloading forklift truck with 18-inch omni wheels was selected for advanced development. This vehicle, the Omni-Directional Ordnance Handler (ODOH), is designed to transport long, heavy missiles and other ordnance down narrow passageways onboard Navy ships. A preliminary design for the ODOH (Figure 5) has been completed.

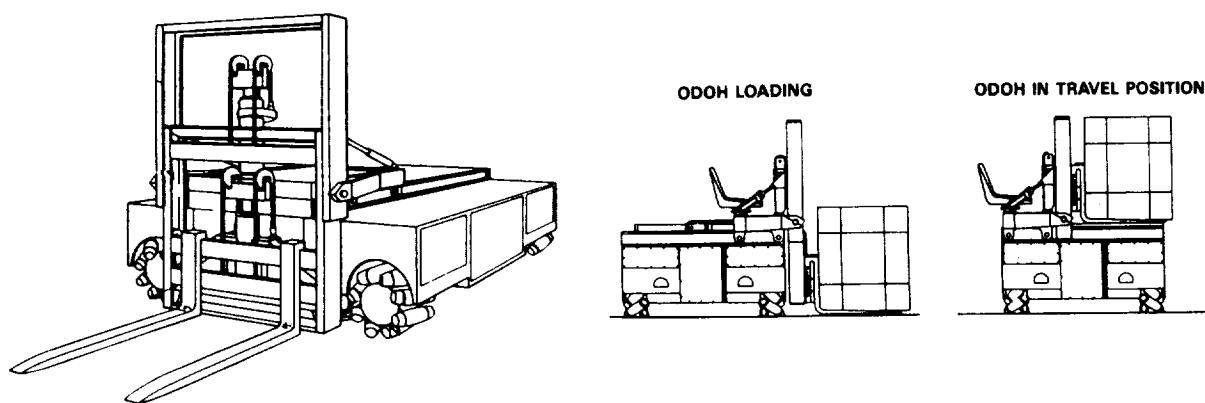


FIGURE 5. OMNI-DIRECTIONAL ORDNANCE HANDLER (ODOH)

A small ODV model (Figure 6) was fabricated to test the electronic control system and to illustrate the omni wheel principle of operation. This model clearly demonstrated the simplicity of operating an ODV, its ability to negotiate obstacles and to maneuver in extremely confined spaces.

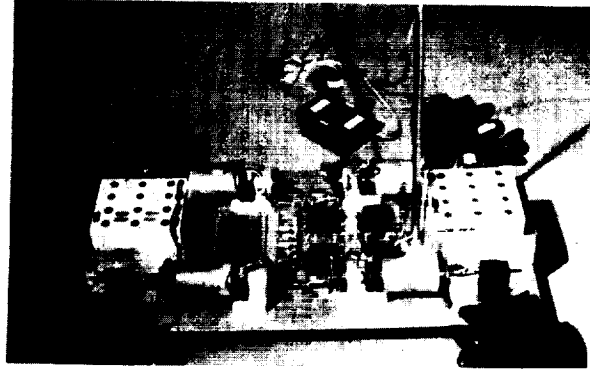
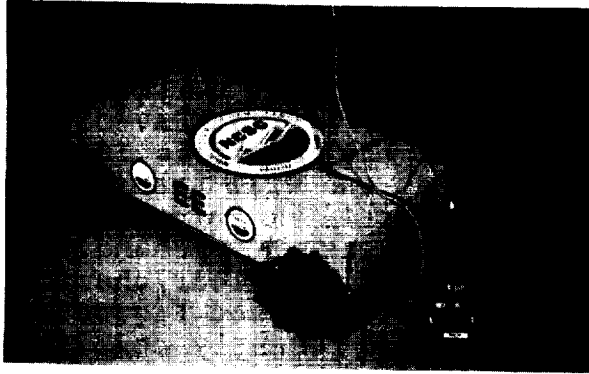


FIGURE 6. OMNI-DIRECTIONAL VEHICLE MODEL

To provide a full-scale Omni-Directional Test Platform (ODTP) (Figures 7a and 7b), the original Cadillac-Gage ODV, (Figure 7c), was modified by: removing the operator seat and other structure, leaving only a simple transport platform; adding an electronic control system to resolve the deficiencies of the original hydraulic system; and replacing the gasoline engine with a battery powered, eight-horsepower electric motor so that the vehicle could be operated indoors and below decks.



FIGURE 7A. OMNI-DIRECTIONAL TEST PLATFORM (ODTP)



FIGURE 7C. CADILLAC-GAGE VEHICLE

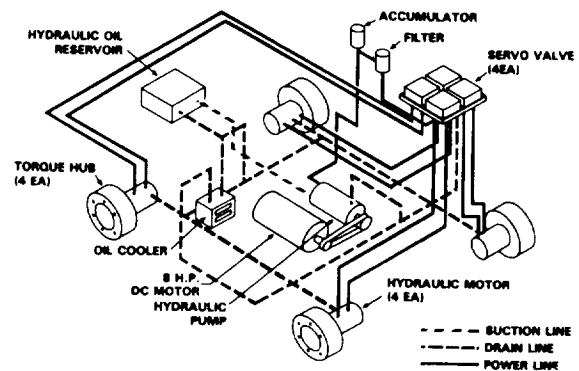
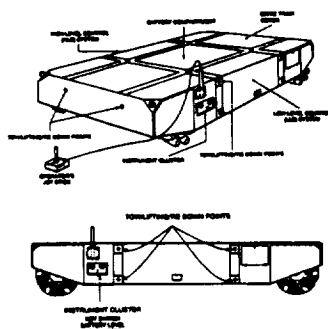


FIGURE 7B. OMNI-DIRECTIONAL TEST PLATFORM (ODTP) DIAGRAM

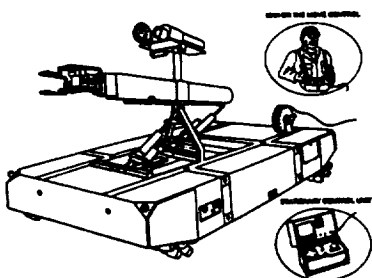
A proof-of-concept Multi-Purpose Autonomous Vehicle (MPAV) Platform (Figure 8) was developed for the Naval Air Warfare Center (NAWC) during 1992-93 to explore the use of ODV technology as a universal platform supporting the use of robotic systems on

aircraft carriers. At this time, the MPAV-ODV is a tethered system but in the future will incorporate high level autonomous control to reduce manpower, improve productivity, and to relieve personnel of hazardous environments and dangerous situations. The MPAV-ODV is 118 inches long, 50 inches wide, 21 inches high and has 18-inch omni wheels. The vehicle weighs 4,700 pounds and will transport a 4,000 pound payload. When equipped with applications hardware, the MPAV-ODV will perform various missions such as cargo handling, weapons loading, jet aircraft engine handling, nuclear/chemical washdown, deck cleaning/deicing, and firefighting (see Figure 9). Shipboard demonstrations of the MPAV-ODV were recently completed and extensive tests of the vehicle with mission adapter hardware are planned for 1995.

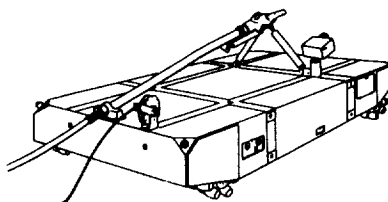


- DIMENSION - 50 INCHES (W), 118 INCHES (L), 21 INCHES (H)
- CAPACITY - 4000 LB PAYLOAD
- VEHICLE WEIGHT - 4700 LBS
- SPEED - 264 FEET / MIN (3.0 MPH)
- RAMP CAPABILITY - 15 DEGS
- OBSTACLE NEGOTIATION - 3 INCH FORWARD, 1.5 INCH LATERAL
- BATTERY POWERED WITH BRUSHLESS DC DRIVE (TRACTION) MOTORS
- ENDURANCE - 8 HRS
- VEHICLE CONTROL - PENDENT OR AUTONOMOUS (HLC) SYSTEM

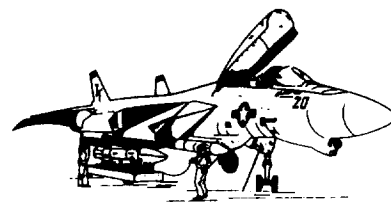
FIGURE 8. MULTI-PURPOSE AUTONOMOUS VEHICLE (MPAV)



HAZMAT PLATFORM



REMOTE FIRE FIGHTING



WEAPON LOADING

FIGURE 9. MPAV-ODV MISSIONS

A small version of the MPAV-ODV known as the Omni-Directional Vehicle, Demonstration Model (ODV-DM) shown in Figure 10 was prepared for NAWC for the development of the High Level Control System and sensors. The ODV-DM, equipped with 12-inch wheels is 32 inches wide, 50 inches long, 21 inches high, weighs 500 pounds, and can transport a 250 pound payload.

OMNI-DIRECTIONAL VEHICLE TEST PROGRAM

The following tests were conducted to demonstrate the capability of the ODTP and other ODVs to operate under demanding conditions:

- a. Vehicle control and traction tests on ice in a skating rink and under cold weather conditions.
- b. Operational tests on a dynamic ship motion simulator.
- c. Static tilt table tests to validate the capability to operate at extreme deck angles.
- d. Missile handling in a simulated shipboard environment to demonstrate the capability of an ODV to transport long, heavy missile loads in restricted spaces.

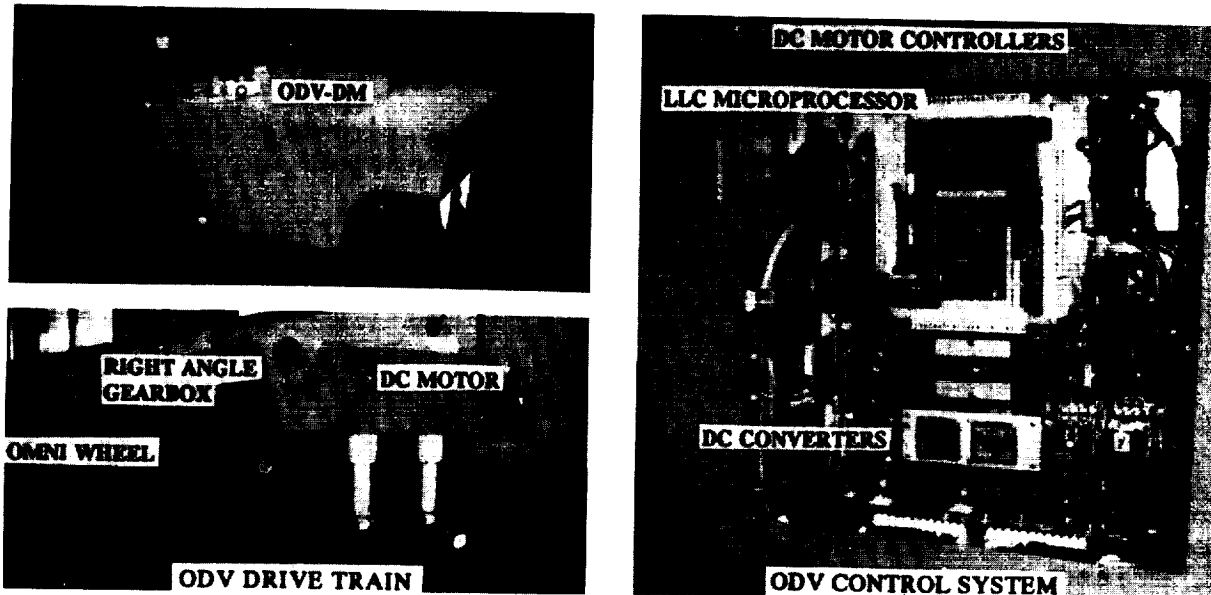


FIGURE 10. OMNI-DIRECTIONAL VEHICLE DEMONSTRATION MODEL (ODV-DM)

These tests have validated the capability of the ODV to operate under adverse shipboard conditions (i.e. reduced traction on wet/icy decks with ship motion and decks heeled at high angles).

Traction Test. The full-scale ODTP, small ODV model, and a conventional forklift truck were operated on ice in a skating rink (Figure 11) to evaluate the capability of an ODV to maneuver and to

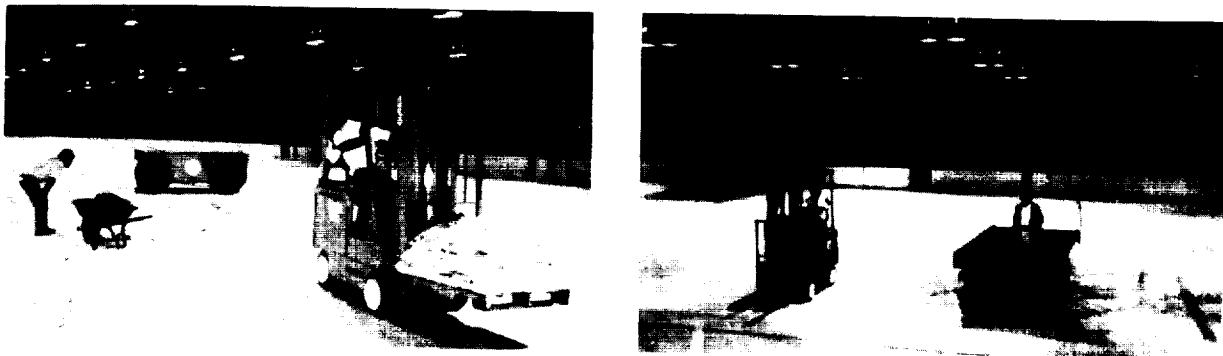


FIGURE 11. ODV ICE RINK TRACTION TESTS

retain control under low traction conditions, and to compare performance with that of a conventional forklift truck. The tests indicated that both ODVs had adequate traction to be fully controllable and capable of performing all maneuvers. The ODTP significantly outperformed the forklift truck in terms of traction and controllability.

Army Cold Weather Traction Test. Tests conducted by the Army further confirmed the ability of the ODV to operate on ice, snow, and wet surfaces. From the test results the following conclusions were drawn:

- a. The omni wheel significantly outperforms a conventional, non-pneumatic tire in driving traction and control on a smooth ice surface.
- b. The omni-directional wheel shows a broad peak traction region on a drawbar-pull versus slip curve, with a desirable slow tapering off of force after the peak value is reached. The fact that it has a broad range of slip levels where peak and near-peak traction occurs makes it very "user friendly" and forgiving to operators.

Ship Motion Test. During simulated ship motion tests the ODTP with a 4,000 pound load was operated forward, backward, sideways, and rotated in place. Motions to five degrees roll and three degrees pitch were induced. Neither the vehicle nor the operator experienced any control problems.

Static Tilt Table Tests. Tilt table tests as required for forklift trucks were conducted (Figure 12) to validate the capability of the ODTP to operate without skidding. The ODTP maintained position without slip under the following extreme conditions of static tilt:

- 26 degrees, ODTP perpendicular to slope
- 23 degrees, ODTP parallel to slope
- 19 degrees, ODTP 45 degrees to slope

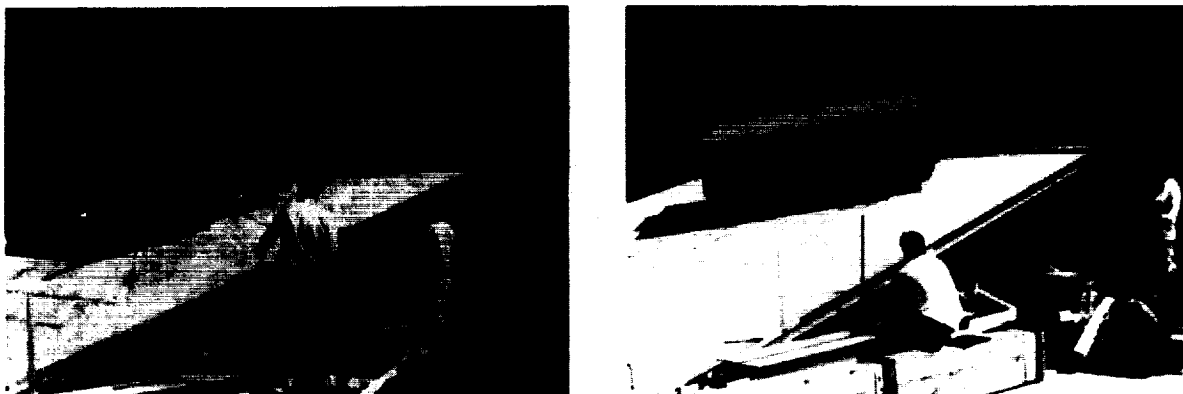


FIGURE 12. ODV TILT TABLE TESTS

Missile Canister Handling. Missile canister handling demonstrations were conducted with the ODTP (Figure 13) to validate the capability of an ODV to transport long, heavy, loads in restricted shipboard spaces. The ODTP performed this mission very efficiently and without difficulty.



FIGURE 13. ODV MISSILE CANISTER HANDLING TEST

MPAV-ODV Functional Test. In September 1992, tests of the NAWC MPAV-ODV were conducted to evaluate vehicle's capability in terms of maneuverability, maintainability, and operability (see Figure 14). Fifteen individuals operated the vehicle after a brief description of the controls. After several minutes of practice, operator proficiency was established and complicated movements were achieved. This observation reinforces one of the primary benefits of the ODV (i.e. the vehicle is capable of complex motions while retaining a simple, user friendly operator interface). This feature is extremely important for man-in-the-loop, teleoperated, or autonomous vehicle operations.



PRECISION MANEUVERING DEMONSTRATION

DRAWBAR PULL MEASUREMENT AT 90°

FIGURE 14. MPAV-ODV FUNCTIONAL TEST



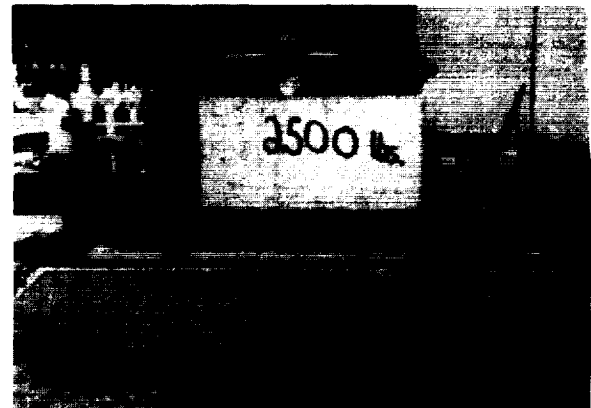
SPEED MEASUREMENT AT 45°



**NEGOTIATING OBSTACLES
(3" BOARD, WIRE, ROPE, CHAIN)**



MANEUVERING LOADED ON 15° RAMP



OVERLOAD TESTS (4900 lbs)

FIGURE 14. MPAV-ODV FUNCTIONAL TEST (Continued)

CONCLUSIONS

Experience with the development of the full-scale ODTP and the recent fabrication and testing of the MPAV-ODV and ODV-DM has demonstrated that ODV technology can be implemented into a practical, rugged vehicle utilizing off-the-shelf components.

As a result of Navy R&D efforts, ODV technology has been validated for applications where maneuverability, control, traction, and obstacle negotiation are required. The ODV is well suited for use in the commercial manufacturer and warehousing environment as well as in nuclear facilities or other hazardous areas.

2515
R 9

APPLYING ROBOTICS TO HAZMAT

Dr. Richard V. Welch

Task Manager, Emergency Response Robotics Project

Mail Stop 138-212

Capt. Gary O. Edmonds

JPL Fire Department, HAZMAT Team Leader

Mail Stop 281-106

Jet Propulsion Laboratory, California Institute of Technology

4800 Oak Grove Drive

Pasadena, CA 91109-8099

ABSTRACT

The use of robotics in situations involving hazardous materials can significantly reduce the risk of human injuries. The Emergency Response Robotics Project, which began in October 1990 at the Jet Propulsion Laboratory, is developing a teleoperated mobile robot allowing HAZMAT (hazardous materials) teams to remotely respond to incidents involving hazardous materials. The current robot, called HAZBOT III, can assist in locating, characterizing, identifying, and mitigating hazardous material incidences without risking entry team personnel. The active involvement of the JPL Fire Department HAZMAT team has been vital in developing a robotic system which enables them to perform remote reconnaissance of a HAZMAT incident site. This paper provides a brief review of the history of the project, discusses the current system in detail, and presents other areas in which robotics can be applied removing people from hazardous environments/operations.

INTRODUCTION

Responding to incidents involving hazardous materials can be extremely dangerous and requires specially trained HAZMAT personnel. Upon arrival to an incident site, the HAZMAT team must first try to determine what types of materials are involved and what threat they present. Unfortunately, records may not be complete or easily accessible and the only way to determine the type and extent of the spill is to send in HAZMAT team personnel.

First entry into incident sites where the types of materials involved have not been identified is particularly dangerous. Members of the team must take all precautions and wear full protective gear including a self contained breathing apparatus and a multi-layer protective suit as shown in Figure 1. This type of protective gear significantly restricts mobility, allows only 15 to 30 minutes of work time, and is extremely hot and stressful on the wearer. Moreover it can take up to an hour for the entry team to suit up once it has arrived at the incident site, delaying identification of the hazard.

The Emergency Response Robotics Project at JPL is prototyping a mobile robot system that can be quickly deployed by HAZMAT teams enabling remote reconnaissance of an incident site without risk to team personnel. The primary goals of the project are:



Figure 1: HAZMAT Team Personnel in Protective Suits

- Develop a teleoperated mobile robot system which can be easily operated by HAZMAT team personnel allowing remote access to an incident site (which may require climbing stairs, unlocking/opening doors, and operating in confined spaces), identification of chemical spills via visual inspection and remote chemical sensing, as well as aid in incident mitigation/containment.
- Work directly with the end-user of such a system (JPL Fire Department HAZMAT team) to establish system requirements as well as use and critique the system under development.
- Work to transfer technology and concepts developed under the project to industry.

These initial goals of the project are discussed in detail in [1]. Other examples of the application of robotics to hazardous material operations are given in [2,3,4].

Several commercially available robotic vehicles were evaluated and two REMOTEC¹ ANDROS Mark V-A systems were procured. (A reference book which covers many of the commercially available and research robots for

¹REMOTEC, 114 Union Valley Road, Oak Ridge, TN 37830

hazardous operations is [5].) The ANDROS robot has a variety of important features needed for the project including its rugged construction, track drive system (enabling stair climbing), manipulator, on-board battery power, and sufficient size to support addition of equipment. Communication between the robot and operator control station is achieved by a 100m tether.

The next section of the paper briefly describes the initial modifications to the ANDROS robot undertaken in the first year of the project leading to the HAZBOT II system. (The name HAZBOT I being given to the "as purchased" system.) The section following this discusses the development of HAZBOT III, a major rebuild of the ANDROS robot. The current status of the project and future plans are then presented. Finally, other areas of potential use of HAZBOT or similar robotic systems are discussed.

HAZBOT II

The most important factor in the development of the HAZBOT II system was training and experimentation with the JPL Fire Department HAZMAT team to determine their requirements. This testing revealed the need for several modifications. One of the most important was the redesign of the operator control panel. The control panel supplied with the system, shown in the bottom of Figure 2, used an array of simple toggle switches to actuate a joint in the robot manipulator. For example, one switch was labeled elbow up/down. This type of control was very difficult for the trainees to master because whether or not the elbow joint caused the forearm of the manipulator to actually move up or down was dependent on the current position or configuration of the manipulator. This type of control therefore led to many mistakes during operation of the manipulator.

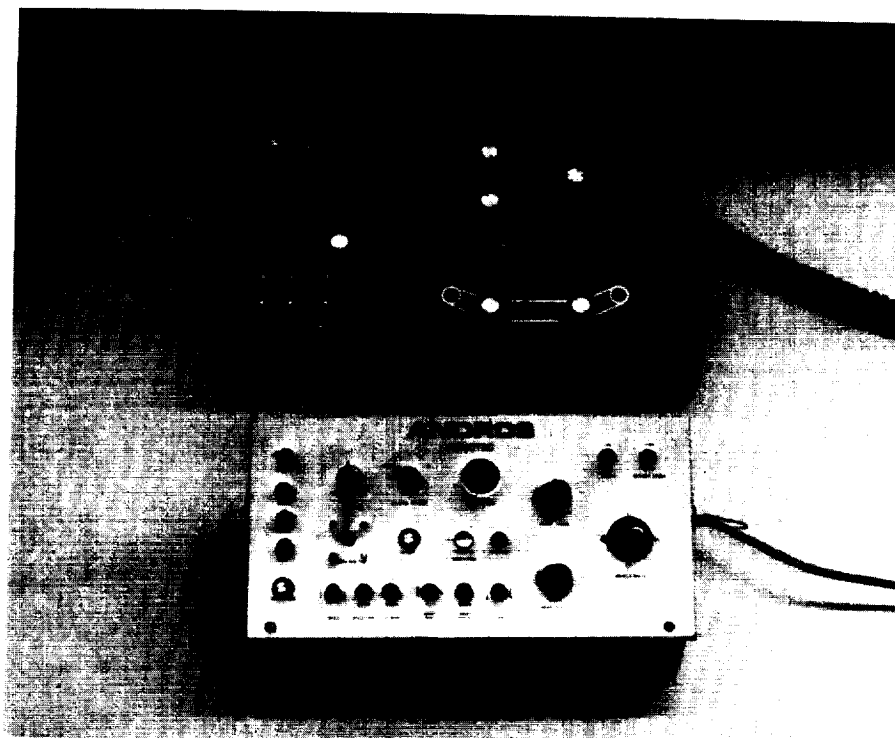


Figure 2: REMOTEC Control Panel and JPL Redesign

A new control panel was constructed that used a simple side view graphic of the robot with controls for each joint placed at the corresponding point of the drawing as shown in the top of Figure 2. The toggle switches were replaced with spring loaded potentiometers; for instance, rotation of the elbow potentiometer clockwise caused the elbow joint to also rotate clockwise. This system was found much more intuitive for the HAZMAT team personnel and led to far fewer mistakes during manipulation tasks.

The HAZBOT II system included a variety of other experimental modifications to the original REMOTEC vehicle such as:

- Development of specialized key tools for unlocking doors.
- Placement of the pan/tilt camera on movable boom allowing better viewing angles during manipulation tasks.
- Addition of a commercial combustible gas sensor often used by HAZMAT teams.
- Addition of a laser depth cuing system.

These modifications are described in greater detail in [6].

We have had active communication with REMOTEC, keeping them up to date on modifications to the system. The control panel redesign has been successfully transferred back to REMOTEC and is being used as a prototype for their new control panels. Currently we are identifying technology in HAZBOT III which can be utilized by REMOTEC in upgrading their own system.

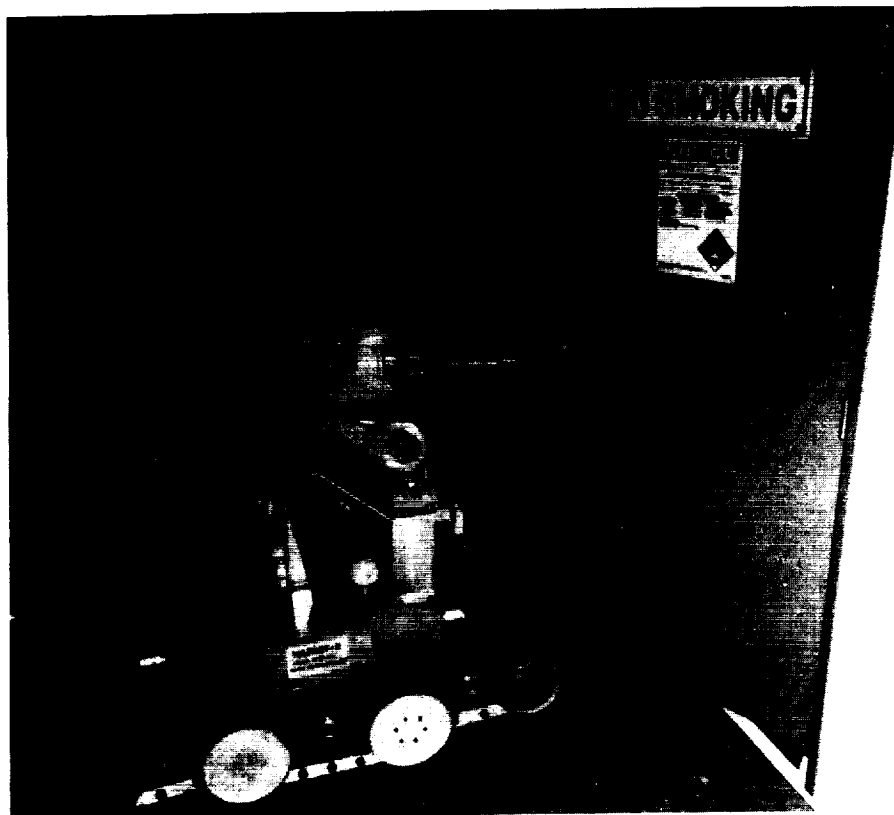


Figure 3: HAZBOT II Unlocking Chemical Storeroom Door

At the end of the first year of the project, a simulated HAZMAT reconnaissance mission was carried out by the JPL HAZMAT team using the HAZBOT II system. The mission (described in [6]) included: opening of the exterior door of a building which had a thumb latch style handle and deployment of a door stop; sensing around a chemical storeroom door for combustible vapors; unlocking and opening of the storeroom door (as shown in Figure 3); and operation in the very small storeroom locating a simulated chemical spill. The operator control station used for the mission, including video displays, the tether spool, and control panel, is shown in Figure 4.



Figure 4: Operator Control Station

Although the use of mobile robots in HAZMAT operations was shown feasible by this first year demonstration, a variety of issues were identified that must be addressed for the system to be used in real response missions:

- Redesign of the robot so that it can operate in environments that may contain combustible gases. This is particularly important in first entry situations where the type of hazard is unknown and potentially combustible.
- Redesign of the robot with a smooth profile and appropriate sealing so that it can be easily decontaminated after a mission.
- Improvement of manipulator in terms of speed and dexterity.
- Continued enhancement of the operator controls.
- Addition of tetherless operation to allow deployment of vehicle greater than 100m from incident site and increase its mobility.

The next section describes how these requirements and the lessons learned in the first year of the project have been used to develop the HAZBOT III system.

HAZBOT III

The focus of the second year of the project was to significantly redesign HAZBOT I (the ANDROS robot that had not been modified in the first year) to meet the system requirements enumerated in the previous section. The primary motivation in design of the new system was the need for operation in Class I, Division 1 environments as defined by the NEC (National Electric Code): environments which contain ignitable concentrations of flammable gases. A two tiered approach was used to address this design requirement. First, all electrical components that may cause electrical arcs or sparks during normal operation were replaced with solid state devices. This included using solid state relays instead of mechanical relays and replacing the brushed DC motors with brushless motors. As a second precaution, all areas of the robot that contain electrical components that could fail and cause sparks are pressurized. The system was not designed to be hermetically sealed but rather to support a small pressure above atmospheric so as not to allow any combustible vapors to enter the system while in operation.

HAZBOT III incorporates the following modifications and features:

- A five foot reach manipulator with a 40 lb payload capacity.
- Parallel jaw gripper with 30 lb grip force.
- Smooth profile to ease decontamination and reduce possibility of snagging during manipulation tasks.
- Internal channels to support pressurization of manipulator.
- Provisions for two movable booms on torso (one currently being used for pan/tilt camera) which also include channels for pressurization.
- A Ross-Hime Designs³ 3 DOF OMNI-Wrist.
- An AIM 3300⁴ specific gas and general combustible gas sensor integrated into forearm and drawing samples in through tip of gripper.
- Use of all brushless DC motors.
- A wrist mounted camera to aid in manipulation tasks.
- Increases of up to 7.5 times in joint speed over original manipulator.
- Low backlash through the use of harmonic drives.
- Reduction of manipulator weight from 150 lbs to 100 lbs.

Other important features of HAZBOT III include a winch system which can be deployed by the manipulator, a microphone and speaker allowing 2-way audio communication, a front mounted tool holder, and an on-board pressure tank.

The chassis of the robot was also enlarged to house a VME type computer system and control electronics. The original ANDROS vehicle used a simple computer system with open loop control of the manipulator. The new VME system includes a 68030 CPU, closed loop control of the new 6 axis manipulator, a variety of analog and digital I/O, as well as room for expansion. Software has been developed using the VxWork⁵ real-time operating system. This computer system provides a solid foundation for future development in coordinated manipulator motion, automation of sub-tasks such as tool retrieval/storage, as well as remote sensing.

³Ross-Hime Designs, Minneapolis, MN 55414

⁴AIM USA, Houston, TX 77272

⁵Wind River Systems, Alameda, CA 94501

In early 1993, HAZBOT III was used to perform a second simulated HAZMAT mission in conjunction with the JPL HAZMAT team. The mission, carried out in the waste material storage facility at JPL, was modeled after an actual incident which had occurred at the site a year earlier. The mission included:

- Unlocking and opening an exterior gate to the facility.
- Locating a simulated spill through an inspection window in storeroom door.
- Unlocking and opening the door to the storeroom as shown in Figure 5 (utilizing the same keytool used to unlock gate).
- Deployment of absorbent pads on spill.
- Opening of cabinet from where the leak was detected.
- Visual inspection and identification of a broken container responsible for spill.

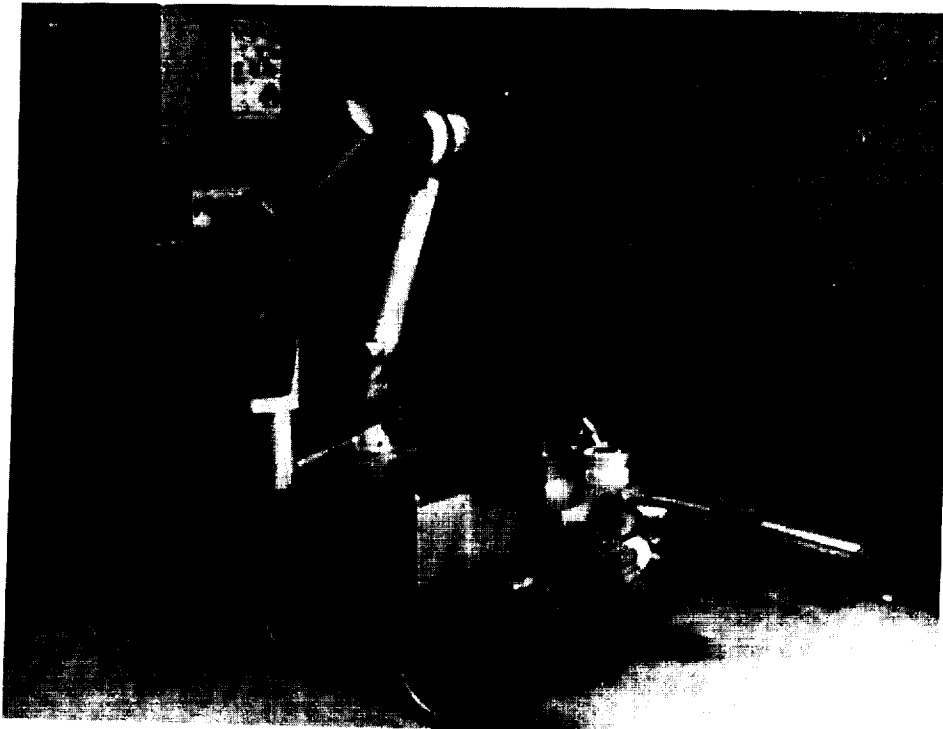


Figure 5: HAZBOT III Unlocking Storeroom Door

Most currently, the track drive sub-system is being upgraded with brushless motors and the pressurization system tested to complete the system rebuild for operation in combustible environments. Training and experimentation of HAZBOT III by the HAZMAT team will continue and help identify areas for continued development. Another simulated response mission is planned for late 1993.

FUTURE PLANS

The HAZBOT III system has addressed many of the requirements as defined by the Fire Department and project team. Two important issues which will be explored over the next year of the project are:

- Tetherless operation - Depending on the type of incident, the robot may have to be deployed at a distance to the incident site greater than its 100m tether length. Also, complex site entry with multiple doors, stairs, etc. increases the chance of snagging the tether and delaying or ending the mission. The tether can be replaced by an RF link for two video signals, 2-way audio, and 2-way data communication.
- Operator controls - The control panel developed in the first year of the project for HAZBOT II has also been used for HAZBOT III. Although a significant improvement from the original design, a wide variety of enhancements can be made to make the operators job easier. (Operator fatigue is a major problem in teleoperations.) These include control algorithm development for coordinated manipulator motion, automation of simple sub-tasks such as tool retrieval/storage, addition of a graphical display indicating system status, sensor data, and vehicle kinematics. (One of the most redundant tasks undertaken by the operator is verification of manipulator position/orientation by scanning with pan/tilt camera.) Additional sensors will also be added to provide information to the operator. It is important to note that the users of this system are not researchers or engineers but Fire Fighters. The controls and feedback to the operator must be in a form that makes sense to them and allows them to confidently use the system for HAZMAT operations.

OTHER APPLICATIONS

Injury or loss of human life can be prevented by using robots in hazardous environments and operations. Robots are now routinely used in industry performing potential dangerous operations such as welding, painting, and material movement. More general purpose robots that can fulfill the need of HAZMAT and other dangerous operations are just crossing the line of economic feasibility. A few years ago, the use of robots by bomb disposal teams was unheard of, while today nearly every major municipal police department has a mobile robot at their disposal. (Newspaper articles describing the exploits of such systems are becoming ever more frequent.) These robots do not replace the highly trained and skilled people in police and fire departments, but rather provides an additional tool that can protect them from injury or death. Other areas for applications of mobile robots similar to the ones discussed here are:

- Mining operations - Not only in general mining operations but perhaps more importantly in gaining access to a mine after an accident. Often the build up of methane or other combustible gases keep rescue teams from entering a mine until it has vented; a system designed for operations in such atmospheres could explore the accident site immediately and help save lives.
- Remote Sampling - Unfortunately today we are faced with many hazardous material dumps which must be monitored on a regular basis. Mobile robots can be stationed at these sites to provide remote sensing and data gathering capabilities rather than repeatedly sending people into the area. Entry into newly discovered sites (for example, those found during military base closures) is very dangerous because the types of materials and the extent of the danger is unknown. Teleoperated robots enable people to remotely and therefore safely explore and classify these sites.
- Law enforcement - As mentioned above, mobile robots are now widely used for bomb disposal. Such systems have also seen duty in hostage situations and armed stand-offs. Robots provide law enforcement agencies remote eyes and ears helping to catch criminals with reduced risk to department personnel.

SUMMARY AND CONCLUSION

This paper has described the Emergency Response Robotics Project and the development of the HAZBOT robots at JPL. The project, currently in its forth year, is prototyping a teleoperated mobile robot for use by the JPL Fire Department HAZMAT team in responding to incidents involving hazardous materials. Key features of the current system include:

- Mobile operator control station with two video displays, custom control panel, and tether reel.
- Tracked mobility system with articulated front and rear sections allowing stairs to be traversed.
- Real-time computer system proving basis for future system development.
- Custom 6 DOF manipulator with integrated chemical gas sensor.
- System designed for operation in combustible atmospheres by using non-arcing electrical components (brushless motors) and internal pressurization.

Two simulated response missions have shown that the basic system is capable of first entry/reconnaissance type missions. Continued system development and training with the HAZMAT team will lead to a robotic system that can be used to respond to actual incidents involving hazardous materials thereby reducing the chance of injury or death of HAZMAT team personnel.

A critical factor in the system development is the close interaction of the project researchers and engineers with the Fire Department HAZMAT team and other safety personnel. This type of directed project and interaction with the end-user or customer must take place if robotics are to move from the laboratory to real-world application. Moreover, industry must be brought into the loop if this technology is going to be made commercially available in a timely fashion.

ACKNOWLEDGMENT

The research described in this paper was carried out at the Jet Propulsion Laboratory, Californian Institute of Technology, under contract a with the National Aeronautics and Space Administration. Special thanks to Henry Stone who managed the first two years of this project, Tim Ohm and Ray Spencer whose technical abilities made the system operational, and the members of the JPL Fire Department HAZMAT team. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

BIBLIOGRAPHY

- [1] H. W. Stone, G. Edmonds, and K. Peterson, "The JPL Emergency Response Robotics Project," *Proceedings of the 1991 JANNAF Safety and Environmental Protection Subcommittee Meeting*, pp. 285-303, Kennedy Space Center, FL, July 1991.
- [2] H. B. Meieran, "Frequent Incidents Increase Robot Potential," *HAZMAT World*, 2(9) pp. 70-73, September 1989.
- [3] H. B. Meieran, "Mobile Robots Responding to Toxic Chemical Spills," *Proceedings of the Forth American Nuclear Society Topical Meeting on Robotics and Remote Systems*, pp. 383-391, Albuquerque, NM, February 1991.
- [4] K. A. Roy, "Mobile Robots Audition For Emergency Response Roles," *HAZMAT World*, 2(9) pp. 64-68, September 1989.
- [5] K. R. Brittain, S. O. Evans, and R. F. Sturkey, "A Compendium of Robotic Equipment Used in Hazardous Environments," *Electric Power Research Institute Technical Report EPRI NP-6697*, February 1990.
- [6] H. W. Stone and G. Edmonds, "HAZBOT: A Hazardous Materials Emergency Response Mobile Robot," *Proceedings of the 1992 IEEE International Conference on Robotics and Automation*, pp. 67-73, Nice, France, May 1992.

ADVANCED TELEOPERATION
Technology Innovations and Applications

Paul S. Schenker, Antal K. Bejczy, and Won S. Kim
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive/ MS 198-219
Pasadena, CA 91109
schenker@telerobotics.jpl.nasa.gov

ABSTRACT

The capability to remotely, robotically perform space assembly, inspection, servicing, and science functions would rapidly expand our presence in space, and the cost efficiency of being there. There is thus considerable interest in developing "telerobotic" technologies, which also have comparably important terrestrial applications to health care, underwater salvage, nuclear waste remediation and other. Such tasks, both space and terrestrial, require both a robot and operator interface that is highly flexible and adaptive, i.e., capable of efficiently working in changing and often casually structured environments. One systems approach to this requirement is to augment traditional teleoperation with computer assists -- *advanced teleoperation*. We have spent a number of years pursuing this approach, and highlight some key technology developments and their potential commercial impact. This paper is an illustrative summary rather than self-contained presentation; for completeness, we include representative technical references to our work which will allow the reader to follow up items of particular interest.

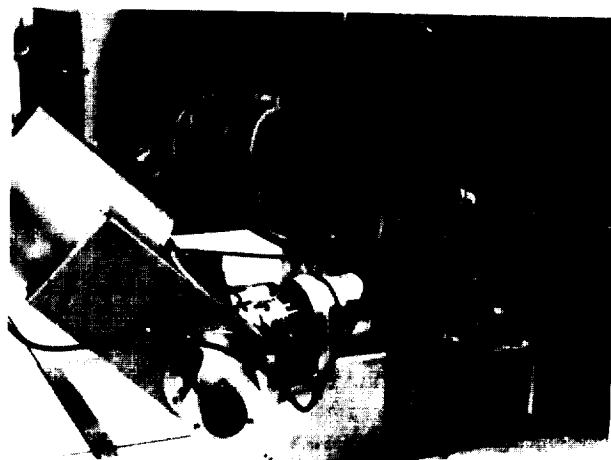
A BRIEF TECHNICAL OVERVIEW

Telerobotics technology development [1] is motivated by a desire to remotely perform complex physical tasks under human supervisory control. To date, robotic systems that have embodied significant supervisory (autonomous) control of their manipulation functions have been limited to highly structured tasks that were performed under favorable and certain conditions -- by definition not complex tasks, and not adaptive performance. This has fostered the widespread use of teleoperation, which at the other extreme from automation, is a characteristically laborious manual control procedure, historically applied to hazardous environments such as nuclear materials handling, underseas recovery, and recently, space shuttle operations. *Virtual environments* and virtual reality (VR) engineering are related and currently popular areas of technology development, wherein the human operator directly manipulates or experiences a modeled, rather than physical reality via computer-synthesis and appropriate input/output devices (e.g., master control gloves/stereo-immersive displays). There exists an important technical intersection of VR technology with telerobotics, most specifically with teleoperation: Virtual environments are useful tools for simulation and design, including task analysis, training, and on-line task preview and prediction. Thus, if VR can be efficiently integrated and physically calibrated with teleoperation systems, it has promise to assist the operator's on-line perception, planning, and control functions.

With regard to space applications, teleoperation systems could have important near-term roles in remote platform servicing, telescience, and lunar exploration, as already illustrated in Shuttle STS-RMS operations. However, the physical and logistical demands of space telemanipulation, particularly in less structured environments, will be high. Tasks can be physically complex and time-consuming, and the operator's manual dexterity and hand-to-eye motion calibration must be good. Further, the work will often be conducted under degraded observational conditions and thus be tedious and fatiguing. Operational uncertainties include obstructed viewing and manipulation, as well as the very disorienting effects of possible time-delay between the operator inputs and robot actions (a major obstacle to achieving desirable ground versus on-orbit operations). In the face of these collective challenges (which have their metaphors in other applications areas such as minimally invasive medical robotics and deep sea teleoperations), we have been trying to "computer-enhance" the performance of traditional teleoperation, and have made progress in the technical areas of redundant telemanipulator control, viewing systems, real-time graphics-based task simulation and predictive control, integrated operator interface design, systems-scale ground laboratory experiments and accompanying human factors data collection & analysis. The laboratory photographs of the next page give a sense of our system implementation; we comment below on our specific enabling technical advances (with supporting citations). For the reader seeking a detailed engineering overview of this work through end-1991, see reference [2]

ADVANCED TELEOPERATION TECHNOLOGY

Validation Through Simulated Satellite Repair Task



A main experimental thrust in our lab has been end-to-end system-level performance characterization -- formal experiment design, integrated system demonstrations, task instrumentation & data capture, and human factors analysis. Collectively, the goal has been to quantify operator limitations, component technology requirements, and their interdependencies in the context of tasks simulated with realistically posed operational constraints (variable lighting, task geometry, time-delay, control & communication bandwidths, viewing & display limitations, etc.). Accompanying human factors issues are the assessment of technology impact on operator error, workload, and training, each in itself a significant risk and cost driver for space operations. As noted above, *advanced teleoperation* is computer-assisted telemanipulation, wherein the operator retains manual control of the task, but with extended functional capabilities and reduced cognitive complexity of task interaction. The computer assists we have developed to date encompass interactive task planning/simulation aids [3], graphics user interfaces for system programming/command/status display [5], and several modes of force-referenced teloperator control which are tolerant to operator positioning error (e.g., "shared compliance control" as described in [2,7] and references therein). In its most general form, advanced teleoperation entails sensory fusion and decentralized control, given that the system sensing, planning, and control functions are inherently distributed between operator and computer; to this end, we have developed some generalized control models and architectures for man-machine interaction at multiple levels of control abstraction, also related multisensor fusion models and techniques [6]. Regarding conventional controls, we have investigated a variety of kinesthetic position, rate, force-feedback, and shared compliance teleoperations modes [2,7]; these controls were first applied to dual six degree of freedom (d.o.f.) PUMA manipulators and more recently to high-dexterity eight d.o.f. redundant manipulators [8], whose controls development has included computer-based techniques of task redundancy management and visualization. We have quantitatively evaluated the operator utility of these these control modes, along with more traditional position and rate approaches, through simulated space servicing experiments [7]. As one example, we performed human factors-based experiments which telerobotically re-enacted high dexterity Solar Maximum Mission satellite repair procedures originally performed by astronaut extra-vehicular activity (EVA) during the 1984 space shuttle flight STS-13. Other supporting developments include real-time graphics environments which allow the operator to animate, analyze, and train on teloperator tasks, and in a most general case, actually use the graphic virtual environment as a basis for reliable teleoperation under multiple second time delay [3,4]. We believe the area of *graphics-augmented reality* for teleoperation has particular promise in space applications and comment further by way of an illustrative example.

AN APPLICATION HIGHLIGHT

A significant obstacle to the acceptance of space telerobotic systems is the impact they might have on operational timelines of crew and platform resources. If a significant part of this burden could be shifted to ground operations, then the technology benefits of space robotics would be far greater. Serendipitously, utilizing ground operations would also free the operator control station of many space-borne implementation constraints, e.g., high degrees of computational power could be brought to bear. However, ground operation of a space robot performing a complex task confronts a basic system limitation in that robotic automation is not yet sufficiently generalized to allow conventional missions control by uplink sequencing of discrete high-level commands. Rather, the operator's continuous direct manual control and eye-to-hand perceptual coordination is required and unfortunately, the implied ground-to-orbit teleoperation approach will not alone suffice either. The problem lies in time-delay communications transit (2-10 seconds latency in current operations scenarios). The operator cannot "fly-by-wire" confidently or coordinate his eye-to-hand skills when causal action-reaction is on the order of seconds; indeed, people rapidly adopt a laborious move-and-wait behavioral pattern when round-loop control latencies are greater than .25 seconds.

Our approach to resolving this fundamental limitation has been to develop a class of 3-D graphics display which visually simulates the robot response in real-time immediacy to the operator's input. In essence, the operator interacts with a virtual task model. Thus, the critical details of the task (and robot itself) must be accurately modeled, and further, must be very accurately geometrically calibrated to the operator's time-delayed visual reality as displayed by down-linked video. In terms of practical implementation, this results in a 3-D high-fidelity graphics display which must be correctly registered and overlaid in translation, scale, and aspect re. a multi-camera robot workspace presentation. See the second page of laboratory photographs for a representative example. Our development of this *predictive graphics display* (with a calibrated virtual reality) has enabled us to preserve the operational features of teleoperation, and reliably operate with intermittent time delays up to 5-10 seconds. In a recent demonstration depicted in the lab photos, we, in coordination with colleagues at NASA Goddard Space Flight Center, performed a

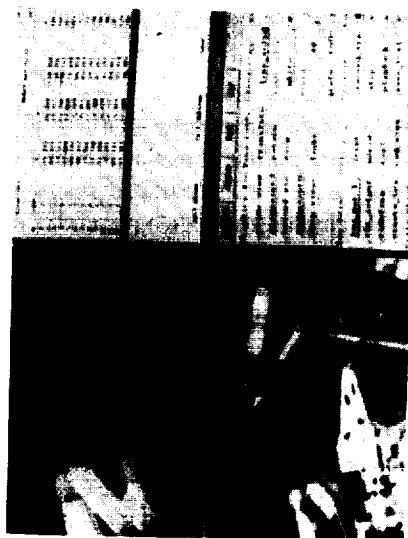
ADVANCED TELEOPERATION WORKSTATION Dual-Arm Control with Graphics Displays for Task Preview and Time-Delayed Operations



CALIBRATED VIRTUAL ENVIRONMENT FOR ADVANCED TELEOPERATION JPL-to-GSFC Time-Delay Operations for Simulated HST Platform Repair



(time delay remote video with calibrated 3-D graphics overlay)



(robot operator's multi-media display during task)

simulated ground-to-remote on-orbit equipment changeout similar to that anticipated for future Hubble Space Telescope servicing: from JPL, having geometrically modeled and visually calibrated the "remote" GSFC robot site, we teleoperatively detached and remounted an ORU. The motion planning and execution, both in free space and guarded-contact, were generated by pure teleoperation, with accuracies of millimeters over a work volume of several meters cubed.

COMMERCIAL MARKETS

The ability to calibrate and animate a virtual environment with respect to actual visual robotic workspaces appears to have significant applications potential. As one example, in the area of medical robotics, it suggests a number of possibilities for computer-guided stereotaxic procedures, microtelerobotic surgery, telesurgery proper (actual remote surgical theatres), also multisensory data presentation and visualization. And of course, calibrated VR seemingly is a key ingredient in planning and executing telerobotic operations in remote scenarios subject to either time delay and or partial viewing obstruction. To this end we have joined with Deneb Robotics, Inc., of Auburn Hills, MI, to cooperatively develop a calibrated 3-D graphics-on-video function within their line of 3-D graphics simulation products.

ACKNOWLEDGEMENTS

This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

- 1) A. K. Bejczy, "Sensors, controls, and man-machine interface for advanced teleoperation," *Science*, Vol. 208, pp. 1327-1335, June 1980; P. S. Schenker, "NASA research & development for space robotics," *IEEE Trans. Aerospace Electr. Sys.*, vol. 24, no. 5, pp. 523-534, September 1988; C. R. Weisbin and M. D. Montemerlo, "NASA's telerobotics research program," *Proc. 1992 IEEE Intl. Conf. on Robotics and Autom.*, Nice, France, May 1992.
- 2.) P. S. Schenker, A. K. Bejczy, W. S. Kim, and S. Lee, "Advanced man-machine interfaces and control architecture for dexterous teleoperations," in *Oceans '91*, Honolulu, HI, October, 1991 (survey paper on JPL Advanced Teleoperation work through fall, 1991, copy available from first author)
- 3.) A. K. Bejczy, W. S. Kim, and S. Venema, "The phantom robot: predictive displays for teleoperation with time delay," *Proc. 1990 IEEE Intl. Conf. Robotics & Autom.*, Cincinnati, OH, May; W. S. Kim and P. S. Schenker, "Teleoperation training simulator with visual and kinesthetic force reality," in *Human Vision, Visual Processing, and Visualization*, *Proc. SPIE 1666*, San Jose, CA, February 1992; W. S. Kim, "Virtual reality calibration for telerobotic servicing," submitted to 1994 IEEE Intl. Conf. Robotics & Autom. (preprint available from authors).
- 4) W. S. Kim, P. S. Schenker, A. K. Bejczy and S. Hayati, "Advanced graphic interfaces for telerobotic servicing and inspection," in *Proc. 1993 IEEE-RSJ Intl. Conf. IROS*, Yokohama, Japan, July; W. S. Kim, P. S. Schenker, A. K. Bejczy, S. Leake, and S. Ollendorf, "An advanced operator interface design with preview/predictive displays for ground-controlled space telerobotic servicing," in *Telemanipulator Technology and Space Robotics*, *Proc. SPIE 2057*, Boston, MA, September, 1993.
- 5) P. Lee, A. K. Bejczy, P. S. Schenker, and B. Hannaford, "Telerobot configuration editor," in *Proc. IEEE Intl. Conf. Systems, Man, and Cybernetics*, Los Angeles, CA, November, 1990; P. Fiorini, A. K. Bejczy, and P. S. Schenker, "Integrated interface for advanced teleoperation," *IEEE Control Systems*, vol. 13, no. 5, pp. 15-20, October, 1993.
- 6) S. Lee, E. Zapata, and P. S. Schenker, "Interactive and cooperative sensing and control for advanced teleoperation," in *Sensor Fusion IV: Control Paradigms and Data Structures*, *Proc. SPIE 1611*, Boston, MA, November 1991; S. Lee, P. S. Schenker, and J. Park, "Sensor-knowledge-command fusion paradigm for man/machine systems," in *Sensor Fusion III: 3-D Perception and Recognition*, *Proc. SPIE 1383*, Boston, MA, November, 1990.
- 7) H. Das, H. Zak, W. S. Kim, A. K. Bejczy, and P. S. Schenker, "Operator performance with alternative manual modes of control," *Presence*, vol. 1, no. 2, pp. 201-218, Spring 1992.
- 8) A. K. Bejczy and Z. F. Szakaly, "An 8-d.o.f. dual arm system for advanced teleoperation performance experiments," in *Proc. SOAR '91 Symposium (Space Operations, Applications, and Research)*, Houston, TX, July 1991; see also, S. Lee and A. K. Bejczy, "Redundant arm kinematic control based on parameterization," in *Proc. 1991 IEEE Intl. Conf. on Robotics and Autom.*, Sacramento, CA, April.

Unit

TEST AND MEASUREMENT

CONTINUOUS MEASUREMENT OF AIRCRAFT WING ICING

Stephen S. C. Yao
Software Manager
Axiomatics Corporation
3G Gill St.
Woburn, MA 01801

ABSTRACT

Ice formation on the wings of aircraft is a problem that has plagued air travel since its inception. Several recent incidents have been attributed to ice formation on the lifting surfaces of wings. This paper describes a SBIR Phase I research effort on the use of small flat dielectric sensors in detecting a layer of ice above the sensor. These sensors are very small, lightweight, and inexpensive. The electronics package that controls the sensor is also small, and could be made even smaller using commonly available miniaturization technologies. Thus, several sensors could be placed on a surface such that a representative ice thickness profile could be measured. The benefits offered by developing this technology go beyond the safety improvements realized by monitoring ice formation on the wings of an aircraft. Continuous monitoring of anti-icing fluid concentrations on the ground would warn the pilot of impending fluid failure as well as allowing the stations to use less de-icing solution per aircraft. This in turn would increase the safety of takeoffs and reduce the overall discharge of de-icing solution into the environment, thus reducing the biohazard of the de-icing procedure.

INTRODUCTION

Several technologies currently exist for detecting ice formation on wings. The older, more established ones use mechanical vibration as the basis for their measurement. These instruments have been proven to be quite bulky, and very difficult to mount on a wing. Newer technologies based on measuring the capacitance of materials above a sensor have shown great promise in measuring ice thickness. All of these technologies measure in-flight icing very well. However, none of them can make accurate measurements in the various ground icing situations.

Axiomatics has applied Shunting Dielectric Sensor (SDS) technology to detecting and analyzing layers of ice on an aircraft surface. This technology measures the complex admittance of materials above the sensor. The admittance is essentially the vector sum of the capacitance and the conductance of the sensor. Using this added information, Axiomatics has been able to simultaneously measure the layer thickness and concentration of solutions.

The coatings on aircraft wings in ground icing conditions could be pure ice, an anti-icing fluid with a variable concentration of water, or an anti-icing fluid/water solution in which ice is starting to form. Since the Axiomatics sensor can measure capacitance, measurements of pure ice layers can be made as well as any of the other capacitive sensors. The advantage of the SDS system is in its ability to make concentration measurements. The concentration of water in anti-icing fluids is thought to be an indication of its ability to function effectively. As the amount of water reaches a certain level ice will begin forming in the solution. The SDS system can be used to monitor the amount of water contained in the solution.

Axiomatics successfully completed a Phase I SBIR contract with the NASA Lewis Research Center in June 1993. The results of this project demonstrated the feasibility of using the SDS sensor in measuring layers of ice, water, de-icing fluid, and mixtures of the three. A \$500,000 Phase II effort has been proposed to begin in the first quarter of 1994. This effort will last for two years, by the end of which Axiomatics will have developed a system ready for flight testing and qualification. Axiomatics is currently seeking partners to participate in this final phase of development.

THE SENSOR SYSTEM

The sensor was designed using Axiomatics' proprietary SDS technology. This technology uses a three terminal flat sensor that provides a distinct advantage over more conventional two terminal sensors. In a two terminal configuration the dielectric measurement relies on the change in energy absorption into the material in question. For many substances the frequency of excitation would have to be in the GHz range before changes in dielectric properties would manifest themselves in absorption changes large enough to be measured.

The SDS introduces a third electrode as shown in Fig. 1. This electrode shunts away some of the electric field from the sensing electrode. This has the effect of increasing the time constant of the sensor and thus lowering the resonant frequency. Using this shunting effect measurement the SDS can provide accurate measurements in the MHz range, thus avoiding the noise problems and high cost of GHz range systems.

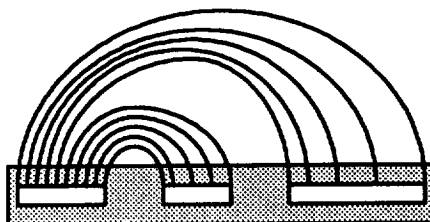


Figure 1: Shunting Dielectric Sensor Electric Field

The field penetration into the material being tested depends on the width and spacing of the electrodes, as well as the dielectric properties of that material. This implies that the amount of shunting will also vary with the same parameters. These facts lead us to an added functionality for the SDS technology, namely, that it is possible, given that the dielectric properties of the material being tested are known, to determine how thick a layer of that material there is above the sensor. The SDS can simultaneously measure both the composition of a material and the thickness.

The sensor is constructed by etching the sensor pattern on a piece of thin film flexible circuit board material. The surface of the sensor is then coated with a tough barrier film to insulate the circuit elements from the environment. This creates a sensor that can be mounted on the exterior of an aircraft with almost no penalty in aerodynamics.

The measurements were made with Axiomatics' Spectrum dielectric measurement instrument. The Spectrum system takes dielectric measurements (capacitance and dielectric loss) of the sensor at ten different frequencies. These frequencies are: 500 Hz, 1 kHz, 5 kHz, 10 kHz, 50 kHz, 100 kHz, 500 kHz, 1 MHz, 5 MHz, and 8 MHz. This dielectric data is then relayed to a computer where it is stored in a disk file.

Temperature measurement of the SDS and sample were obtained using a surface-mounted K-type thermocouple, located just under the sensor. The Spectrum system is currently not able to measure temperatures below 0°C, so a Fluke thermocouple meter was used to make those measurements. Side by side tests of the thermocouple meter and the Spectrum system indicated that they read temperature within 0.5°C of each other for temperatures above 0°C. This is within normal thermocouple accuracy specifications.

ICE/WATER THICKNESS

Testing & Results

Testing was done on layers of ice/water ranging from 0.5 mm to 4 mm. These layers were deposited above the sensor. Measurements were made at temperatures varying from -20°C to 20°C. From the testing we found that there was a dramatic change in capacitance between a layer of ice over the sensor and a layer of water. Figure 2 graphs the capacitance measured by the sensor versus the temperature. The measurement frequency for this graph is 1 MHz. At temperatures below 0°C the

capacitance of the ice layer varies from about 0.45pF to about 0.55pF. There is also a slight linear variation in the capacitance as the temperature varies. At temperatures above 0°C the capacitance of the water layer varies from about 4pF to about 5.5pF. Here there is a somewhat larger variation in capacitance with temperature. At 0°C, as expected, there is a vertical line connecting the ice measurements with the water measurements. Obviously, the variation in capacitance is due to the changing amounts of water and ice in the layer above the sensor.

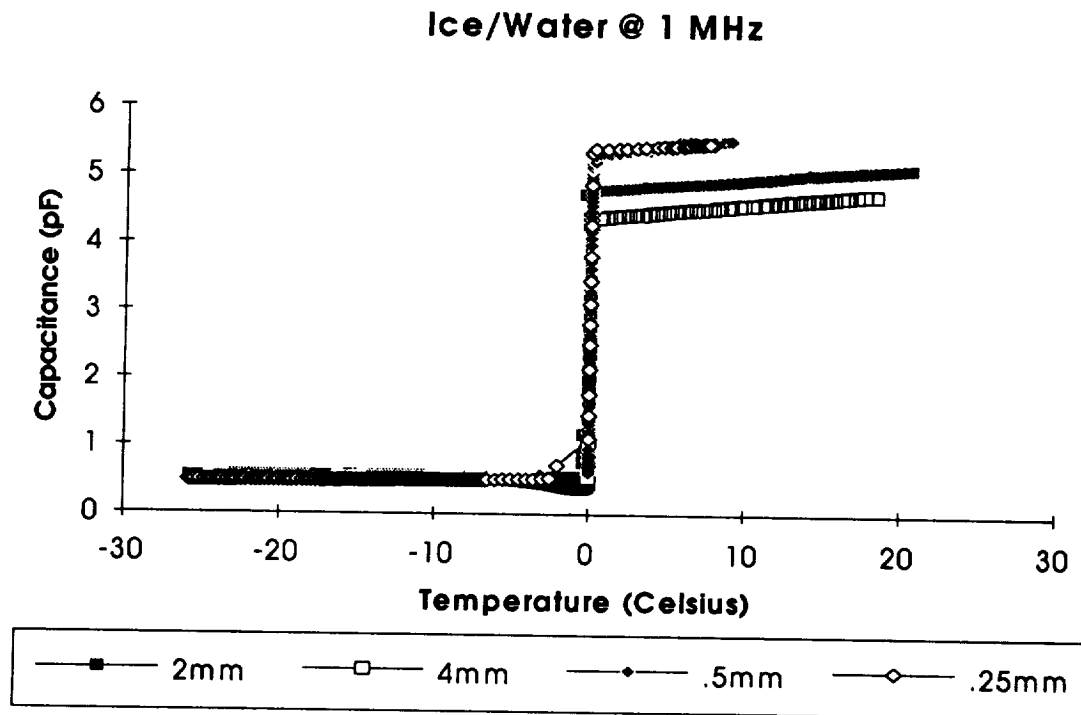


Figure 2

Figure 3 shows the capacitance versus the loss of the ice portion of the testing, i.e., the data taken when there is only ice above the sensor. This graph is essentially the same as a graph in the complex admittance plane (z-plane) with the loss being the real component and the capacitance being the imaginary component divided by 2π times the measurement frequency, in this case 1 MHz. Along with the ice data is data taken with an empty sensor. The capacitance varies almost linearly with the loss as the temperature changes. Also, the slope of this variation seems to be constant for all thicknesses of the ice layer. This is a very interesting result because it implies that it is not necessary to measure the temperature of the ice in order to measure its thickness. In fact, it might even be possible to measure the temperature using only the dielectric information. This may be useful, as it is sometimes difficult to make an accurate measurement of the bulk temperature of the material with a surface mounted thermocouple.

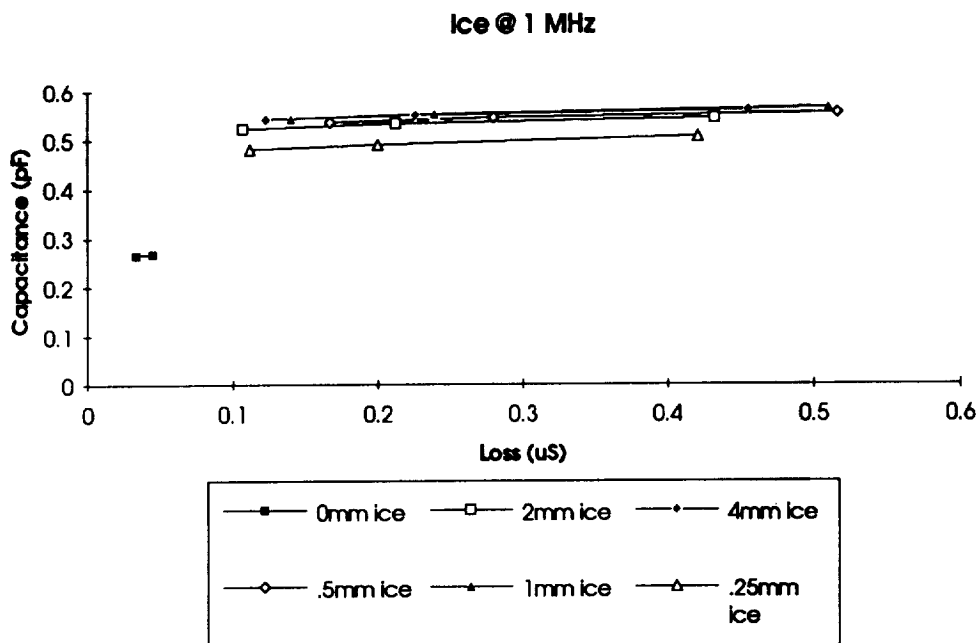


Figure 3

From the data illustrated in the above graph it can be theorized that, as the thickness of the ice layer increases, the capacitance also increases. The thicknesses listed on the graphs are only approximate. It was not possible to make precise measurements with the equipment available for the testing. Even so, the graph shows a definite trend of increasing capacitance until the 0.5mm thickness is reached. At that point all of the thicker layers seem to exhibit the same dielectric response, to within acceptable experimental error. This would imply that the electric field generated by the sensor penetrates 0.5mm into ice.

Ice Thickness Prediction Algorithm

The results of the testing seem to show that all of the data with ice only exhibited a capacitance of 1 pF or lower at 1 MHz. All of the data with water or de-icing fluid exhibited a capacitance of 3 pF or higher. The first step in the algorithm, then, is to check the capacitance at 1 MHz. If the capacitance is below 1 pF but above the C_0 of the sensor, i.e., the capacitance of the sensor in air, then it would be inferred that there is a layer of ice on the sensor.

The second step in the algorithm is to normalize the temperature effects on the ice measurements. This is done by looking at the loss and capacitance in the z-plane. From the results it can be shown that the dielectric values vary linearly with temperature in the z-plane. Not only this, but all of the slopes for the various thicknesses are equal to 0.065. To normalize temperature it only requires that we extend a line through the measured point in the z-plane with the slope equal to 0.065 and determine the capacitance intercept. This will give a temperature normalized capacitance that can then be related directly to ice thickness.

The equation for temperature normalizing the capacitance at 1 MHz is

$$\text{Normalized Capacitance} = \text{Capacitance} - [0.065(\text{Loss})]$$

The relationship of thickness to normalized capacitance at 1 MHz is shown in Figure 9. The prediction equations is

$$\text{Thickness} = 1.54902 \times 10^{-8} e^{33.8207(\text{Capacitance})}$$

To verify these equations, a test similar to the data collection tests was run. The data was taken at one or two temperatures for each layer thickness. The prediction algorithm yields results that predict the thickness of the ice layer within about 0.1mm. Even this small error would probably be considerably

reduced if more data were taken and a better fitting equation were developed. The errors are due more to insufficient modeling points than to inaccuracies in the measurements.

It should be noted that errors exist in the actual thickness. It is very difficult to measure this thickness because of several factors. The surface of the layer is not smooth. Also, for the 0.1mm layer it is not certain whether the entire sensor surface was covered or not. There might be air bubbles trapped in the layer. For the purposes of this feasibility test, however, the accuracy of the measurement was deemed sufficient.

DE-ICING FLUID THICKNESS

Testing & Results

Testing was done on layers of de-icing fluid ranging from 1 mm to 4 mm. These layers were deposited above the sensor. Measurements were made at temperatures varying from -20°C to 20°C . The capacitance data at 1 MHz of varying thicknesses of de-icing fluid shows that the capacitance of layers of de-icing fluid 1mm deep or greater is about 4pF and higher. This would mean that it is quite easy to differentiate between ice over the sensor and thicknesses of de-icing fluid of 1mm or more. In fact, it is reasonable to assume that it would be difficult to create a layer of de-icing fluid over the sensor that would mimic the response of ice.

Figure 4 shows a z-plane graph of data taken at 100 kHz. There seems to be a regular behavior as the temperature changes. The behavior follows the theoretical SDS spiral shape as the temperature changes. This indicates that the dielectric properties of the de-icing fluid change with temperature, which allows us to normalize the dielectric readings according to the temperature. Thus, it should be possible to develop an equation to predict the thickness of a layer of de-icing fluid from the dielectric measurements.

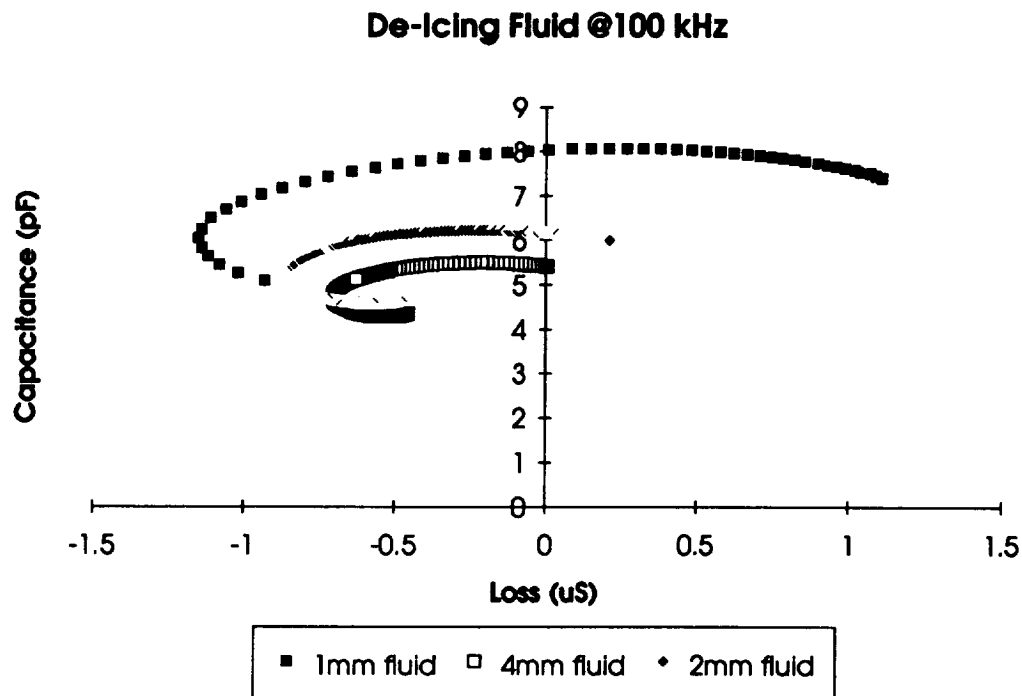


Figure 4

DE-ICING FLUID CONCENTRATION

Testing & Results

Tests were run to determine the response of the sensor to variations in the concentration of de-icing fluid. Solutions were prepared of distilled water in Type II de-icing fluid ranging in concentrations from 0% to 75% water. A thickness of 2 mm of each of these solutions was placed over the sensor. Measurements were made at a temperature of -10°C . Figure 5 shows a graph in the z-plane of the results of this testing. The graph suggests that this sensor system can be used to measure the concentration of the de-icing fluid.

SDS Response to Water in Type II De-Icing Fluid

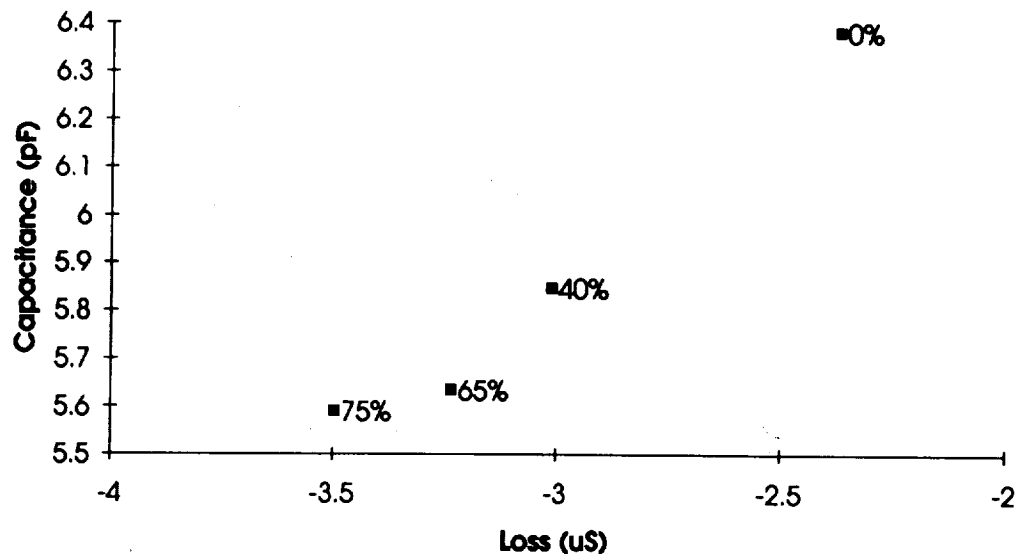


Figure 5

CONCLUSIONS

Axiomatics has achieved the goal of establishing the feasibility of measuring various thicknesses of ice and differentiating ice from water and de-icing fluid under conditions which simulate those of aircraft icing. This conclusion is reached on the basis of the results of the testing done on water, ice, and de-icing fluid. Ice had a very different dielectric response from water and from de-icing fluid, so it was very easy to determine whether a layer was composed of ice or not. The feasibility of measuring the thickness of layers of water or de-icing fluid was also demonstrated.

The accuracy to which the thickness of a layer of ice can be determined, and the maximum measurable thickness of that layer, depend on the geometry of the sensor. In this project the sensor used could measure ice thicknesses up to 0.5mm to within about 0.1mm. The data indicates that improvements could be made if a better method of measuring the thickness of the layer were available. If thicker layers or thinner layers are to be measured then the geometry of the sensor will be changed. It may also be possible to achieve the same results by changing the construction materials.

In addition, the feasibility of measuring the concentration of de-icing fluid has also been shown. It is believed that the effectiveness of de-icing fluid is linked to the amount of water in solution. Having a continuous measurement of de-icing fluid concentration could provide early warning of its impending failure.

**Electromagnetic Probe Technique
For Fluid Flow Measurements**

G.D. Arndt
NASA/Johnson Space Center
Houston, Texas 77058

J. R. Carl
Lockheed Engineering and Sciences Company
Houston, Texas 77058

Abstract

The probes described herein, in various configurations, permit the measurement of the volume fraction of two or more fluids flowing through a pipe. Each probe measures the instantaneous relative dielectric constant of the fluid in immediate proximity. As long as separation of the relative dielectric constants of each fluid is possible, several or even many fluids can be measured in the same flow stream. By using multiple probes, the velocity of each fluid can generally be determined as well as the distribution of each constituent in the pipe. The values are determined by statistical computation. There are many potential applications for probes of this type in industry and government. Possible NASA applications include measurements of helium/hydrazine flow during rocket tests at White Sands, liquid/gas flow in hydrogen or oxygen lines in Orbiter engines, and liquid/gaseous Freon flow in zero gravity tests with the KS135 aircraft at JSC. Much interest has been shown recently by the oil industry. In this a good method is needed to measure the fractions of oil, water, and natural gas flowing in a pipeline and the velocity of each. This particular problem involves an extension of what has been developed to date and our plans to solve this problem will be discussed herein.

Introduction

The development of a microwave technique for measuring two-phase flow was originally started due to a desire to monitor the flow of monomethyl hydrazine and helium through an inlet pipe during tests at the White Sands Facility of a reaction control system (RCS) thruster jet. The relative amounts of helium and hydrazine flowing into the thruster jet could not be measured on an instantaneous basis. It was realized that since the dielectric constants of helium (approximately 1) and hydrazine (19.2) are sufficiently different, the load impedance seen by a microwave capacitance probe should also be sufficiently different to be easily separable.

The microwave technique that is described in this paper measures the phase angle of the reflection factor, S_{11} , associated with reflected energy from a flush-mounted probe. The system is in the process of being modified to include multiple probes within the pipe. This system has other potential space applications in measuring the flow of liquid and gaseous oxygen or hydrogen under zero-gravity conditions within the Space Station. The technique also has ground-based applications in measuring gas-water-oil flow from undersea oil wells as well as other possible uses in measuring volume fractions and the velocity of multiple liquids having different dielectric constants.

Applications

Single Non-intrusive Probe

There are many potential applications for a single, non-intrusive probe. For example, a single probe mounted at the top of a pipe can perform well as a bubble detector or void detector. A single probe mounted at the bottom of a pipe could be used to continuously monitor the purity of the fluid. A single probe mounted at an appropriate position on a mixing tank could monitor a change from fluid A to fluid B as a function of time. A single probe mounted strategically could be used to identify laminar or turbulent flow. A single probe may be all that is needed to monitor some point of interest in a pipeline. Combined with apriori information, flow regimes may be indentifiable using a single probe. Of course, a single probe could be used to identify a full or empty tank, or an intermediate threshold level.

Multiple Non-intrusive Probes

Multiple non-intrusive probes could do any of the things mentioned previously. Identifying flow regimes, and calculating volume fractions could probably be accomplished better with multiple probes located at different positions on the pipe and performing additional processing. Velocity computations would require at least two probes at a known downstream spacing.

Multiple Intrusive Probes

In order to monitor directly what is happening in the interior of a pipe or reservoir, intrusive probes must be used (if using the type probe discussed in this paper). If multiple interior locations are to be monitored, then multiple probes are required. An example of this type of requirement comes from the oil industry. They have a need to measure the volume fraction of oil, water and natural gas flowing through a pipe and the velocity of each. It may be necessary, in this case, to gather data at the interior of the pipe.

If probes are internal, they must be made as minimally intrusive as possible. Also, they must be rigid enough to withstand the flow and tough enough to withstand corrosion and abrasion for long periods of time.

System Description

The major components of a single probe system are shown in Figure 1. This system has been built and used in a test program as described later.

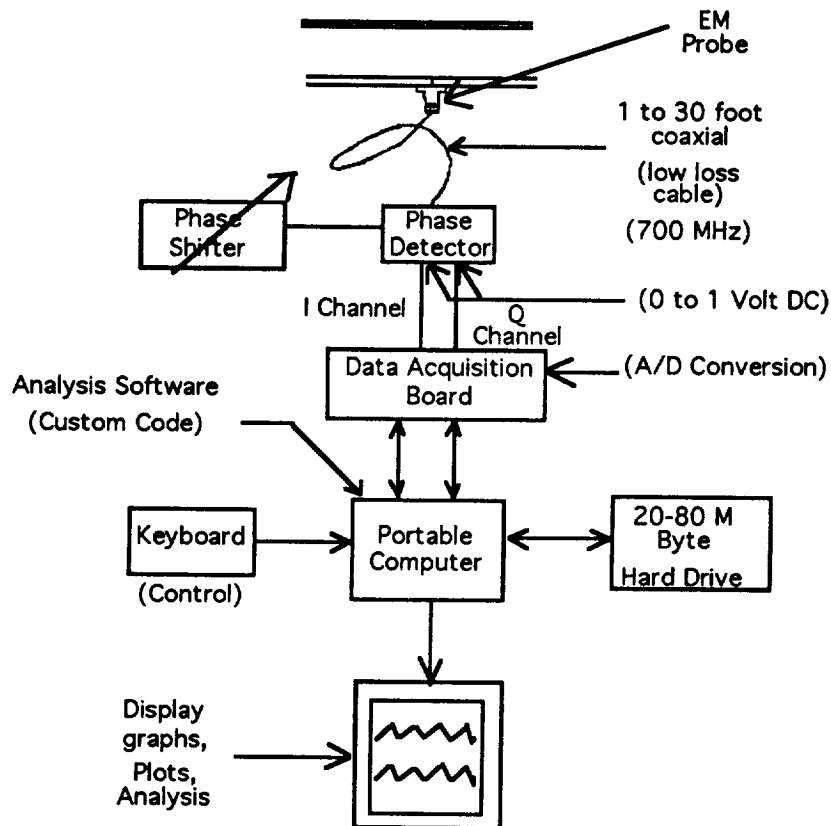


Figure 1. EM Probe System Block Diagram

Probe

The original probe is nothing more than a bulkhead SMA coaxial connector with the center conductor cutoff flush or nearly flush with the inside of the pipe wall. The teflon insulator around the center conductor is cut off flush with the inside of the pipe. Figure 2a shows four such probes mounted around the circumference of a pipe so that it can be determined what is flowing at the top, bottom, and both sides of a pipe. Figure 2b shows two probes at some known spacing to provide the velocity of the fluids flowing at the top of the pipe.

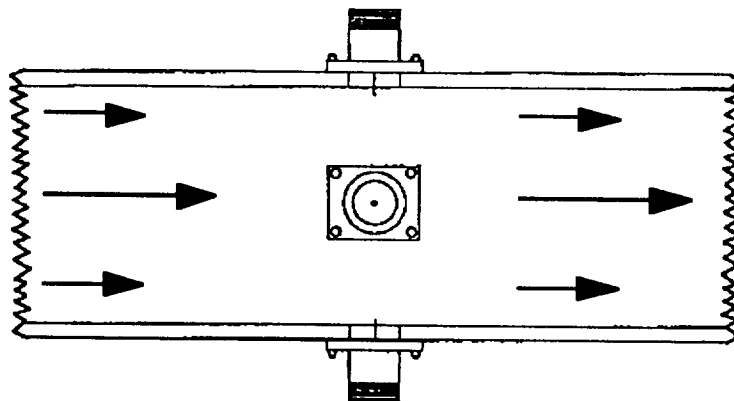


Figure 2a. Multiple Circumferential Probes

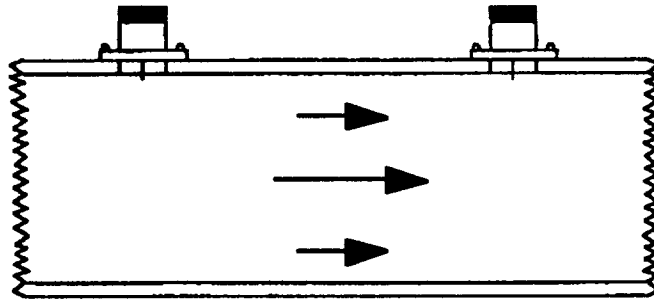


Figure 2b. Multiple Downstream Probes

Phase Detector

A block diagram of the phase detector is provided in Figure 3. This device measures the phase angle on the reflected signal from the probe at approximately 1 GHz. The signal is converted to 100 MHz, amplified, and quadrature phase detected. The two outputs of the phase detector is adjusted to be in the range of 0 to 5 volts. The phase detector was built in-house.

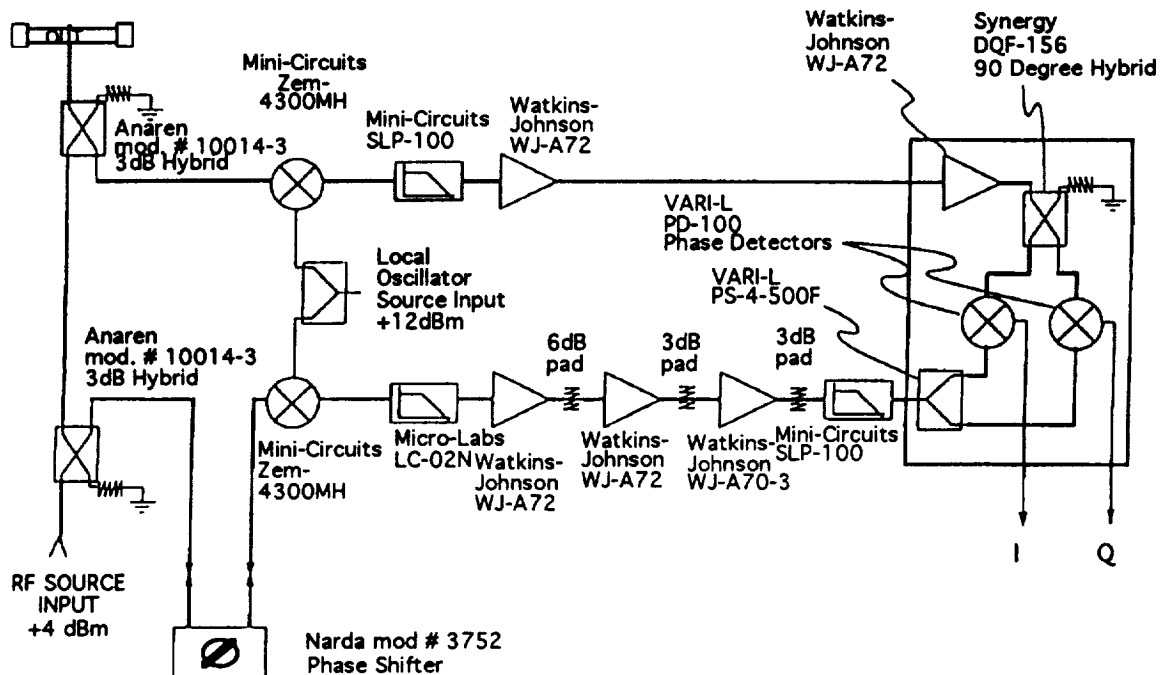


Figure 3. Phase Detector Block Diagram

Data Acquisition Board

There are many commercially available data acquisition boards in the form of expansion boards that are easily inserted into most personal computers (except some of the compact portables and/or notebook computers). Many channels can easily be obtained at reasonable cost and A/D conversion rates are generally available at up to 1 MHz. Conversion rates of 100 KHz are usually adequate.

Portable Computer

A 286 computer with a 20MB hard drive was used for the development of the single probe and used for the testing done to date. However, a 486-DX is being used for the development of the multiple probes with a 120MB hard drive.

Software

The software used for the development and testing of the single probe was written in the "ASYST" language. A commercially available software package called "Viewdac" was purchased for the development of multiple-probe configurations. This package should make it easier to acquire, store, process and display multiple data streams.

Theory of Operation

The small capacitance of the probe is used as the sensor. This capacitance is a function of the relative dielectric constant of the medium into which the probe is terminated. The short probe "sees" fluid that is no more than a short distance away. If it is required to see further into the medium, it would be necessary to increase the length of the probe so as to increase the volume around the center conductor that forms the probe capacitance.

The phase, or change of phase, associated with the reflected signal at the probe is the quantity measured. The complex "S" parameter associated with reflected energy "S₁₁" is given by:

$$S_{11} = \left(\frac{Z_0 - Z_L}{Z_0 + Z_L} \right) \quad (1)$$

Where: Z_0 is the characteristic impedance of the transmission line from phase detector to probe.

Z_L is the probe input impedance = $R + jX$

Typically, Z_0 is equal to 50 ohms. If there is a negligible energy coupled to the media, the probe resistance is very small. The input impedance is essentially a capacitive reactance in which case equation (1) can be written:

$$S_{11} = \left(\frac{50 - jX}{50 + jX} \right) \quad (2)$$

Where: $X = \frac{-1}{2\pi F C \epsilon_r}$

C = probe capacitance

ϵ_r = relative dielectric constant

F = frequency of operation

The phase " ϕ " on S_{11} , from equation (2), can be extracted as:

$$\phi = \tan^{-1} \left(\frac{100X}{2500 - X^2} \right) \quad (3)$$

for certain cases of interest, where X is large with respect to Z_0 and where ϕ is small, equation (3) reduces approximately to:

$$\begin{aligned} \phi &\approx -100 / X \text{ radians} \\ \phi &\approx -5730 / X \text{ degrees} \end{aligned} \quad (4)$$

For a certain 1mm probe, the probe capacitance has been measured to have a capacitance of approximately 0.04 pf. Using this value and using a frequency of 700 MHz equation (4) reduces to the convenient form:

$$\Delta\phi \approx \Delta\epsilon_r \quad (5)$$

For this model of the probe, i.e., a capacitive reactance termination for the transmission line, predictions can be made for probe capacitance given frequency, probe length, and the effective relative dielectric constant of the media. Also the probe's sensitivity can be readily formulated.

Test Results

The results shown in Figure 4 were extracted from the test results of a "Flapper Valve Experiment" performed at NASA, JSC, Houston Texas in November of 1992. The flush probe was mounted at the top of a 1 1/2 inch pipe and monitored the flow of distilled water and dry nitrogen flowing at various specific rates through the pipe. The volume fraction of water and nitrogen were varied. The top graph shows that the probe works well as a bubble detector in this configuration. The lower two plots show two different conditions of slug flow. The precise volume fractions and flow rates are not immediately evident from these plots but by processing the data, introducing apriori knowledge, and by influencing the calculation with calibration data, perhaps reasonably accurate volume fractions and flow rates could be determined using only a single probe.

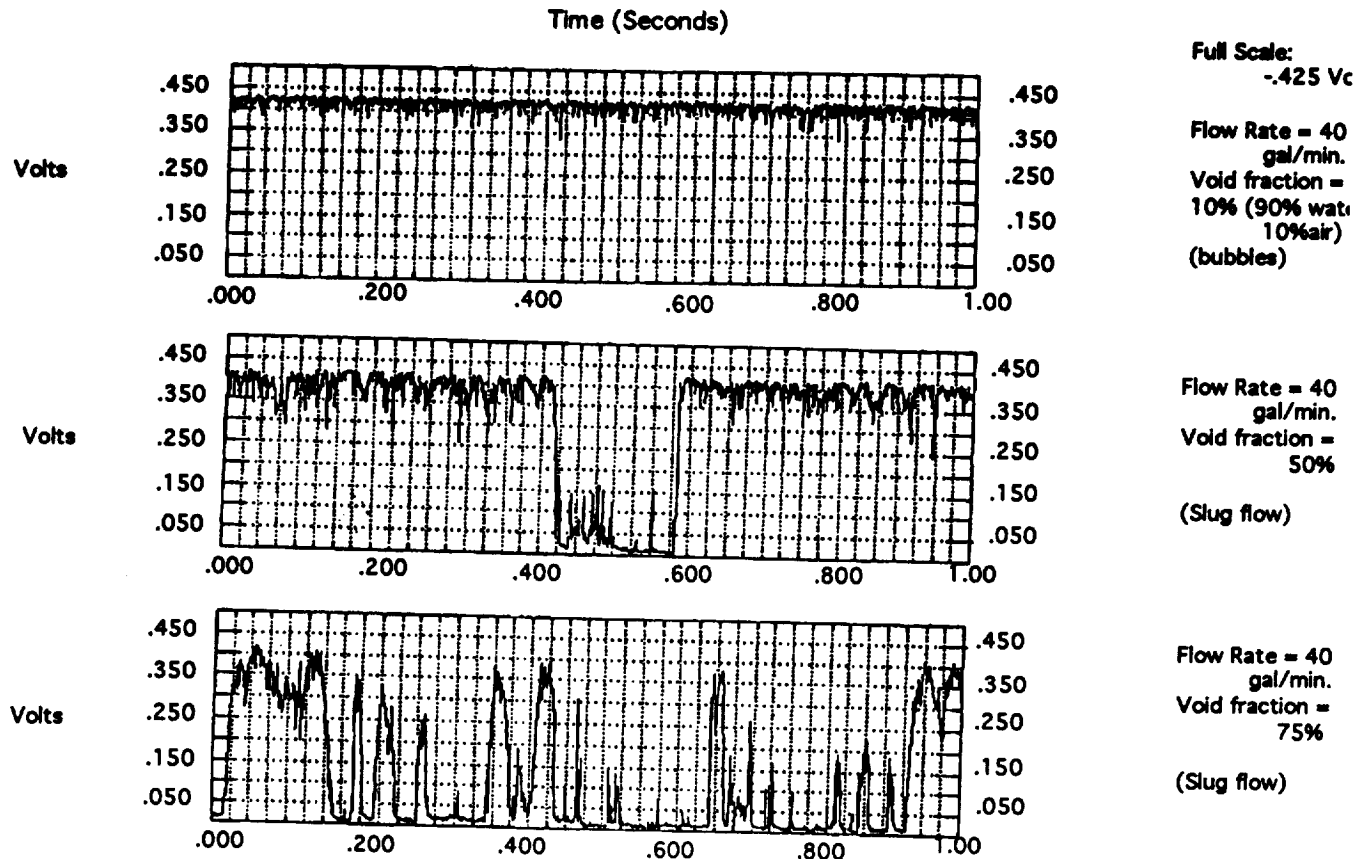


Figure 4. Flapper Valve Experiment Using Distilled Water and Nitrogen

Work In Progress

Certain potential applications of the AC fluid flow probe requires the use of multiple probes placed in the flow stream. For example, consider the problem of measuring the volume fraction of oil, water, natural gas, and oil/water emulsions (flowing in a pipe line). Various flow regimes are possible which may influence the choice of probe locations in the flow stream. By placing probes at various strategic locations within a cross-section of the pipe, the volume fractions of each constituent can be calculated statistically by a technique described herein. Also, by using an identical probe configuration downstream from the first, the velocity of each constituent can, in most cases, be measured.

Multiple in-flow probes can be used to measure many parameters, such as:

- Volume Fractions of Each Constituent
- Velocity of Each Constituent
- Blob Statistics
- Flow Regimes
- Flow Profiles

An Example Configuration of Multiple Probes in the Flow Cross-Section

One probe configuration presently under consideration consists of 2 vertical columns of 7 probes in each column. This configuration is presently being evaluated by computer simulation and is shown in Figure 5. It is believed that, in most cases, a good measurement of gas volume fractions and constituent velocities can be made for the horizontal flow of a non-homogeneous mixture of natural gas, oil, and water. Of course many other probe configurations are possible and can be tailored to the problem. For the flow conditions shown, the gas has a tendency to move toward the center of the pipe.

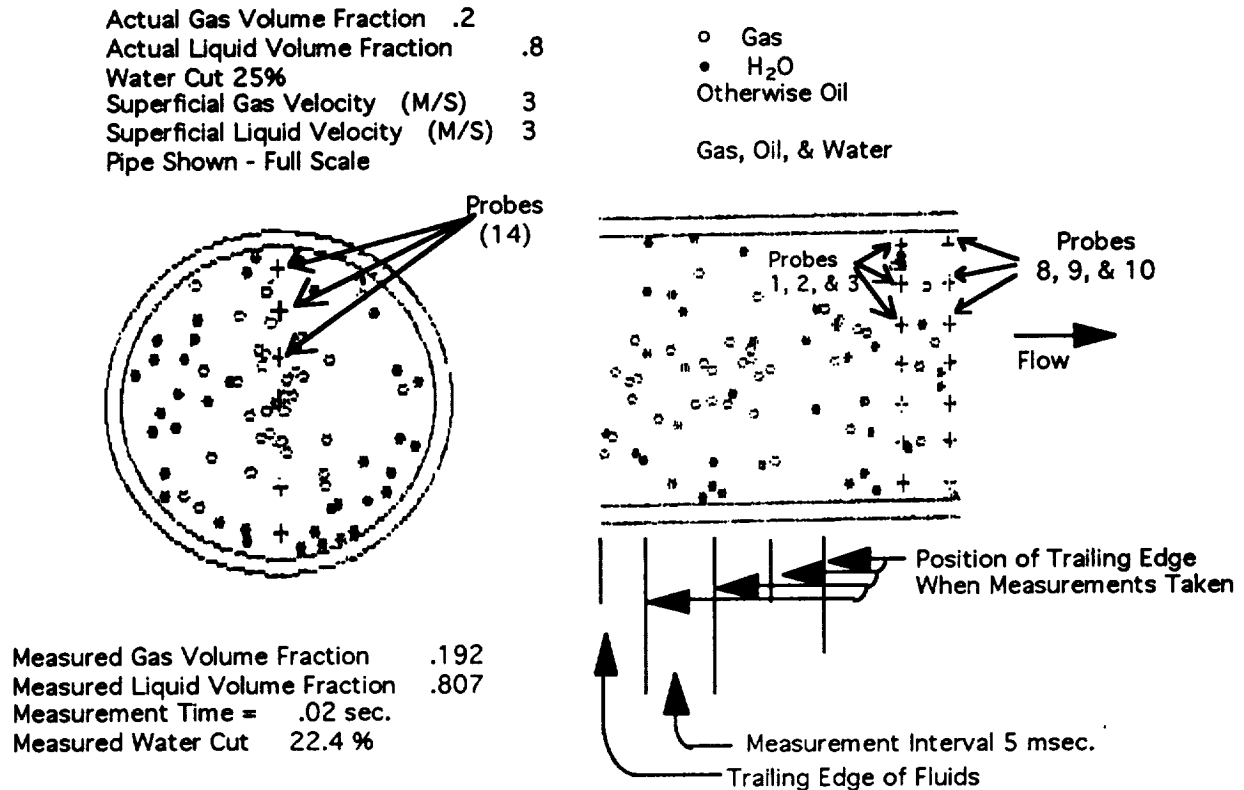


Figure 5. Computer Simulation Results

Multiplexing Probes

It is planned to multiplex the probes so that only one phase detector is required. In order to accomplish this, multiplexing using a switch must be done at RF and the switching device should be placed near the probes. Also the phase detector should be placed reasonably close to the multiplexer to reduce RF losses, maximize the probe sensitivity, and minimize cable induced phase shifts. The output of the phase detector is also multiplexed to sample and hold circuits that are assigned to specific probes. If this multiplexing scheme proves to be unsatisfactory for some reason, multiple phase shifters can be used instead.

Probe Pairs Used to Measure Velocity

The nature of the data streams from each probe is shown in Figure 6. The first column identifies the probe making the measurement. The remainder of the boxes contain numbers that identify the fluid that covers a specific probe at a specific time. Time increases from left to right. Probe 1 and Probe 8 form a an "upstream/downstream pair". Similarly probe 2 and 9 form a "pair" and so on. By sliding the second data stream along the first, a best match can be found. The time displacement required to obtain a match, along with the known probe spacing is all that is required to determine velocity.

This cross-correlation procedure can be performed on all seven pairs and for all three fluids. When this procedure was followed using much longer data streams, velocity could be determined very accurately.

1		1	1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	1	1	1	3	3
2		1	1	1	1	1	1	2	2	2	2	2	1	1	1	1	3	3	1	1	3	3
3		3	3	1	1	1	1	1	2	2	2	2	2	1	3	3	3	3	3	3	1	3
4		3	3	3	1	1	1	1	2	2	2	2	1	1	3	3	3	3	3	3	3	3
5		3	3	3	1	1	1	1	3	3	2	2	1	1	3	3	3	3	3	3	3	3
6		3	3	3	1	1	3	3	3	3	3	3	1	1	1	1	3	3	3	3	3	1
7		1	3	1	1	1	3	3	3	3	3	1	1	1	1	3	3	3	3	3	1	1
8		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
9		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
10		1	1	1	1	1	1	1	1	1	1	3	3	3	1	1	1	1	2	2	2	1
11		1	1	1	1	1	1	1	1	1	3	3	3	3	3	1	1	1	2	2	2	2
12		1	1	1	1	1	1	1	1	1	3	3	3	3	3	1	1	1	1	3	3	2
13		1	1	1	1	1	1	1	1	1	1	1	3	3	3	1	1	3	3	3	3	3
14		1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	3	3	3	3	2

USING PROBE 1 AND 8, THE MEASURED GAS VELOCITY IS 0 M/S

USING PROBE 2 AND 9, THE MEASURED GAS VELOCITY IS 0 M/S

USING PROBE 3 AND 10, THE MEASURED GAS VELOCITY IS 3.29 M/S

USING PROBE 4 AND 11, THE MEASURED GAS VELOCITY IS 3.66 M/S

USING PROBE 5 AND 12, THE MEASURED GAS VELOCITY IS 2.99 M/S

USING PROBE 6 AND 13, THE MEASURED GAS VELOCITY IS 2.99 M/S

USING PROBE 7 AND 14, THE MEASURED GAS VELOCITY IS 2.99 M/S

LEGEND: 1=OIL, 2=WATER, 3=GAS

Figure 6. Data Streams of Each Probe

Multiple Probe Design

Probe structures that are placed in the flow steam must be designed to have minimum impedance to flow and minimum effect on the flow while having good strength and durability characteristics. A prototype probe stack of 3 probes is presently being designed and is shown in Figure 7.

These probes should be large enough to provide a reasonable capacitance at 100 to 300 MHz (for sensitivity), yet small enough to be relatively invisible to the flow. Also the probes should be self-cleaning and should not interfere with each other electrically or physically.

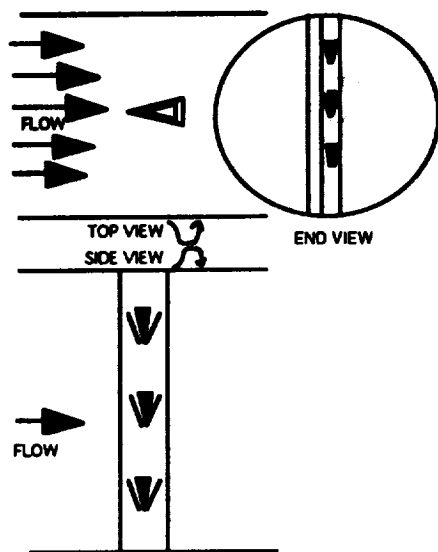


Figure 7. Prototype Stack of 3 Probes

References

A. Cartellier and J.L. Achord, *Local Phase Detection Probes in Fluid/Fluid Two-Phase Flows*, American Institute of Physics, February 1991.

D. Brown, J.J. den Boer and G. Washington, *A Multi-Capacitor Multiphase Flowmeter for Slugging Flow*, Shell Research Paper 2.2, North Sea Flow Measurement Workshop, October 1992.

**COMPUTERIZED ULTRASONIC TESTING SYSTEM (CUTS)
FOR IN-PROCESS THICKNESS DETERMINATION**

J. Frankel, M. Doxbeck, S.C Schroeder and A. Abbate

**U.S. Army Armament, Munitions, and Chemical Command
Army Research, Development and Engineering Center
Benet Laboratories, Watervliet, N.Y. 12189-4050**

ABSTRACT

A Computerized Ultrasonic Testing System (CUTS) was developed to measure, in real-time, the rate of deposition and thickness of chromium plated on the inside of thick steel tubes. The measurements are made from the outside of the tubes with the ultrasonic pulse-echo technique. The resolution of the system is 2.5 μm . (0.0001 in.) and the accuracy is better than 10 μm (0.0004 in.). The thickness is measured using six transducers mounted at different locations on the tube. In addition, two transducers are mounted on two reference standards, thereby allowing the system to be continuously calibrated. The tube temperature varies during the process, thus the input from eight thermocouples, located at the measurement sites, is used to calculate and compensate for the change in return time of the ultrasonic echo due to the temperature dependence of the sound velocity.

CUTS is applicable to any commercial process where real-time change of thickness of a sample has to be known, with the advantage of facilitating increased efficiency and of improving process control.

INTRODUCTION

During processes such as electro-polishing or electro-plating, no real-time information was heretofore available regarding the rate of the process or the thickness of the film. The final thickness was obtained only after the process finished and the specimen cooled, cleaned and dried. For large steel tubes, one or two days are lost before results can be obtained. The choice of proper plating parameters is thus made based only on past experience, usually obtained by means of experimental tests at varying flow and current parameters [1]. This makes the characterization of a plating process cumbersome and extremely expensive, and limits the range of conditions simulated. In the vessel plating technique for large hollow tubes, the bore contains the plating solution, which is pumped in at the bottom at about 85 °C (185 °F) and exits the top at a higher temperature. Currents of the order of 20,000 A are sent through the tube (the cathode), pass through the plating solution and return through the lead-plated anode. [2] (Fig. 1).

A Computerized Ultrasonic Testing System (CUTS) has been designed and developed, which enables the user to evaluate the thickness and the rate of application of chromium films during vessel plating of steel tubes [3]. This system provides continuous information on thickness, rate of deposition and temperature at six different location on the tube, with a thickness resolution of 2.5 μm . (0.0001 in.) and an accuracy of better than 10 μm . (0.0004 in.).

The requirements on thickness accuracy demand extremely stable coupling between the transducer and the tube, and the presence of temperature fluctuations imposes the need for temperature compensation in the equations which calculate the thickness. Due to the high temperatures on the outside surface of the tube, the transducers must be cooled in order to prevent degradation. This will also extend the life span of the transducers. The system incorporates the use of two standards, to provide continuous thickness calibration and drift correction during the measurement.

Computer technology is thus combined with ultrasonics to provide a real time process monitor, automated data gathering, computation and display of the thickness at six transducer locations placed in two

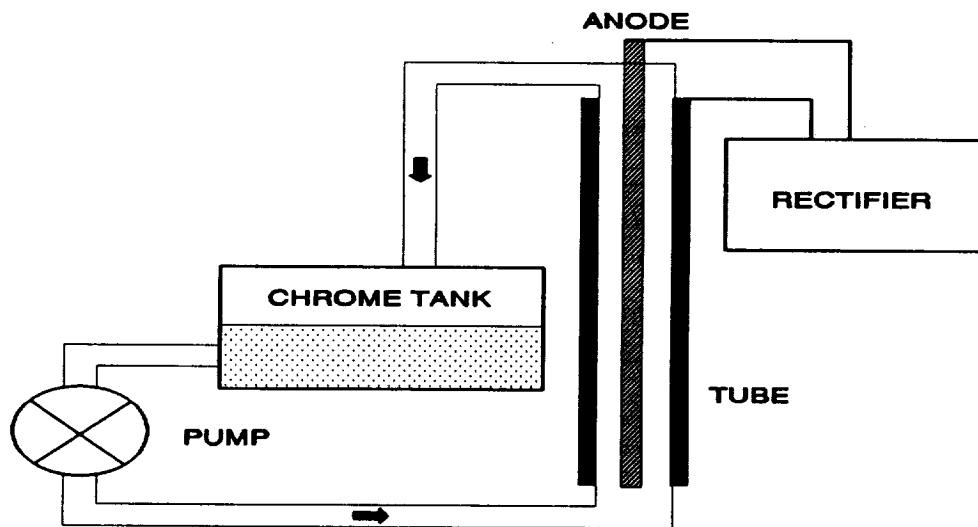


Figure 1 Schematic representation of the flow-through process for large hollow tubes.

rings. [4]

Even though this system was designed for utilization in a vessel plating facility, its principles are general in nature and can be applied to any process in which the thickness of the sample is modified. The fixtures used to couple the sound wave to the sample under study may differ from the one presented here, but the electronics and the software remain essentially the same [3].

PRACTICAL CONSIDERATIONS FOR ACCURATE REAL-TIME ULTRASONIC TESTING

Several concerns were addressed during the design of CUTS and in particular of the transducer assembly: [4]

- The fixtures used to hold the transducers had to be stable enough not to vibrate in the shop environment so that minute changes in the return time, of the order of nsec., could be associated with changes in chromium thickness;
- During plating, the temperature at the outside surface of the steel tube can be as high as 95 °C. A cooling system had to be designed, in order not to damage the transducers;
- In order to measure the distance between parallel surfaces, the ultrasonic beam had to be directed along the radius of the tube;
- The transducer had to be electrically insulated from the plating circuits.

The most feasible solution to these problems was found in using focussed transducers, coupled to the tube through a constant liquid path, with constant alignment such that the beam would travel radially to the tube. (Fig. 2) Commercially available, 5 MHz longitudinal transducers of 0.75 inches in diameter were utilized. Six transducers were placed 120 degrees apart in two rings at two different locations along the axis of the tube. A water cooling line around the cylinder containing the liquid coupling agent was also used to reduce the temperature at the transducer. The configuration of the six transducers was chosen to optimize the information obtained for an overall description of the plating. By comparing measurements between transducers on the same ring, the wall variation of the chromium film around the section could be evaluated. Furthermore, by comparing results between the two rings, the unevenness of plating along the length of the tube and any possible tapering could be estimated. The set-up time of both rings is roughly 30 min and less time is needed for breakdown.

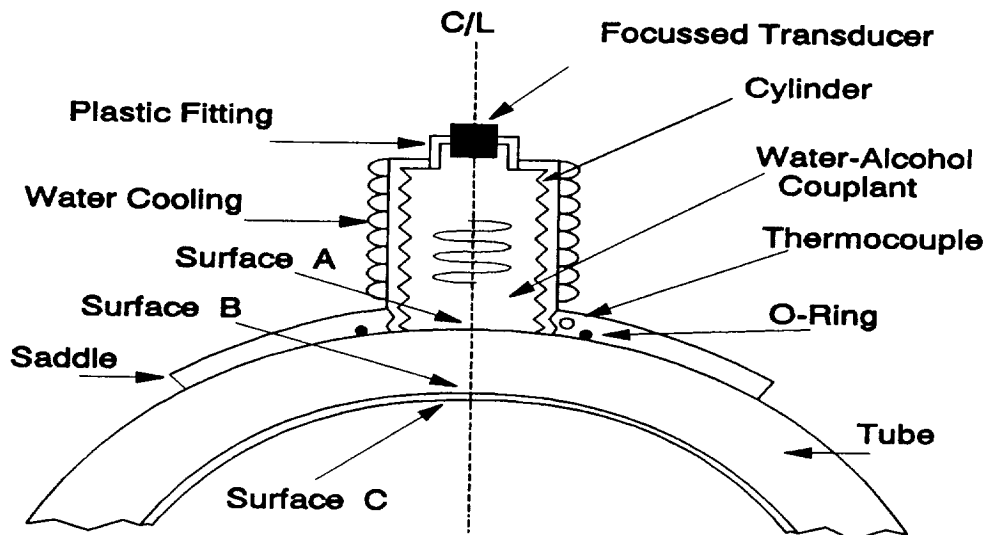


Figure 2 Sketch of the transducer assembly which utilizes liquid-buffer coupling.

The liquid used for coupling is a water-ethylene glycol mixture. When water was used as the couplant, as the cooling coils started to cool the cylinder, additional ultrasonic echoes appeared between A and C in Fig 2. It is surmised that the cooling of the water caused a layer of cooler temperature near the edge. Since the slope of the velocity-temperature curve for water is steep, this caused its velocity and hence acoustic impedance to increase and act to efficiently scatter sound waves at the layer interface. For these reasons and after trials of several mixtures, the 20 volume % ethylene glycol-water mixture was chosen. This mixture exhibits an almost flat velocity-temperature curve peaking around 60 °C, and its use eliminated the anomalous reflections.

The thickness of the sample is obtained by measuring the time necessary for a sound pulse to travel through the specimen and reflect from surface C (Fig. 2). The sound wave, generated by the transducer, travels through the liquid couplant and is partially reflected from surface A. The portion of sound wave which is transmitted through surface A, reaches surface C and then is reflected back. All reflection are thus detected by the transducer at the other end of the path with different time delays related to the different distances covered. If the velocity of sound in the material is precisely known, then the thickness can be calculated by measuring the difference in time between the two reflections. It has to be pointed out that this approach is valid since the sound wave is not reflected at surface B, (the steel-chromium interface).

The sound velocity in the steel tube is measured using two steel samples of known thickness and the temperature at the surface of the tube near the transducer and at the steel standards is monitored by thermocouples and recorded by the computer in order to compensate for the difference in sound velocity. Since the velocity calibration is continuously performed during the process, these measurements are also used to check and correct for any possible zero or baseline drift of the electronic equipment.

DESCRIPTION OF THE SYSTEM

A schematic representation of the measuring system is given in fig. 3. The system consists of a commercially available ultrasonic thickness gage, an oscilloscope, and an IBM-compatible AT/PC with an A/D temperature board. The Panametrics 5215-1C ultrasonic thickness gage, with a resolution of 2.5 μm (0.0001 in.) and dual Automatic Gain Control, was used to measure the time interval between the interface echo (Surface A) and the first echo from the inside diameter (Surface C). This unit has a multiplexer which can be connected to 8 transducers. Gate adjustments allow the time interval measurement to be made for

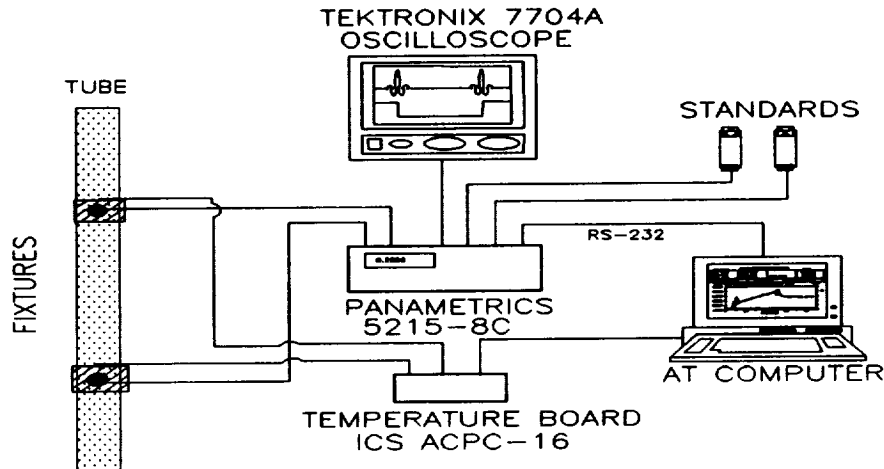


Figure 3 Schematic representation of the ultrasonic thickness measuring system.

any echo combination. Of the eight transducers, six are located on the tube and two on the standards, and the result of the time measurement are transmitted to the computer in ASCII format via an RS-232 port. The steel standards have thicknesses of 2.28727 cm (0.9005 in.) and 1.77800 cm (0.7000 in.).

A software program was written for the acquisition, analysis and display of the results. The program provides a variety of graphs while plating is in progress. One such screen, from an actual plating run, is displayed in Fig. 4. During the run, eight transducers and eight thermocouples were used, though for simplicity, the numerical display in Fig. 4 shows only three transducers. STAND1 and STAND2 represent the readings on the two steel standards and TRANS3 measures the tube thickness. Two markers, defined as START and STOP, are used to evaluate the plating rate. As shown in Fig. 4, the markers have been positioned at time 45 min., (start of electro-plating) and at time 220 min. (end of electro-plate), so that a straight line joining them fits the actual increase in thickness. Here the plating rate is shown to be 59.2 $\mu\text{m/hr}$ (.002331 in./hr) and the thickness difference between the two markers is 172.7 μm . (0.0068 in.). Mechanical measurements performed by our Quality Assurance (Q/A) technical personnel, showed an increase in thickness of $177.8 \pm 18 \mu\text{m}$. (0.0070 ± 0.0008 in.). The three thermocouple readings measure the temperature of the standards (TEMP1), of one of the transducers on the tube (TEMP2) and of the steel near the transducer (TEMP3). The computer cycles in a loop of data acquisition/analysis as long as desired. The data are saved in a ASCII file, and can be imported into any program, such as plotting utilities or spreadsheets, for further analysis.

The Panametrics equipment provides a voltage V proportional to the echo return time τ between the reflection of the sound wave from the surfaces A and C in fig. 2. This time τ can be expressed as the sum of various terms:

$$\tau = \frac{2x_s}{v_s} + \frac{2x_c}{v_c} + \tau_x \quad (1)$$

where x_s , x_c are the thickness of the steel and of the chromium film, respectively; v_s and v_c are the sound velocity in the respective media; τ_x represents any systematic error involved in the measurement. The velocity of sound v_s and the error τ_x are continuously determined using the two steel standards. The difference in temperature between the standards and the sample is also considered in the calculations for v_s . From laboratory testing the velocity of the chromium film was estimated to be approximately 1.13 times the velocity of sound in steel ($v_c = \gamma v_s$), and a similar dependence in temperature is expected for both

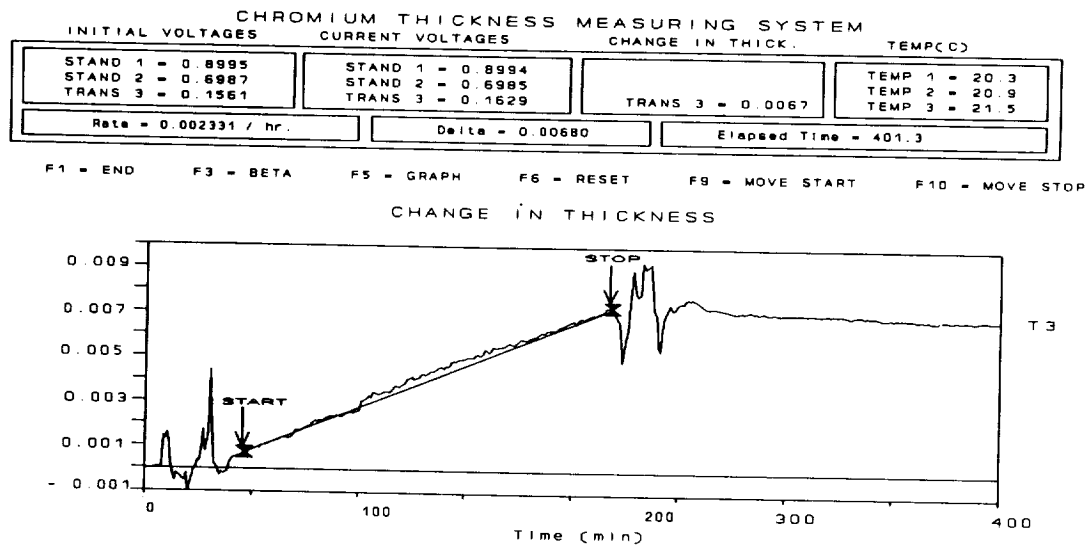


Figure 4 Computer display during the plating run. The straight line is a fit used to obtain the plating rate.

materials.

The change in thickness is evaluated by subtracting the initial value measured before the process started from the absolute measurement of the thickness. In the case of electro-plating, at time zero $x_c = 0$, and the thickness of the chromium film is evaluated by the following relationship:

$$x_c = \gamma \cdot \left[v_s(T) \frac{\tau - \tau_x}{2} - x_s \right] \quad (2)$$

ANALYSIS OF THE SYSTEM

In order to optimize the accuracy of the measurements, an error analysis was performed on experimental data obtained during plating. The standard deviations and thus the errors involved in the evaluation of the chromium thickness x_c are shown in Table I. The total error in the estimation of x_c is 9.41 μm . (0.00037 inches). The largest contribution is given by the uncertainty of the value of the sound velocity in chromium. From eq. 2 it is clear that the effect of γ on the value of x_c is proportional, hence an error of γ of 3%, corresponds to a 3% error in x_c . The film thicknesses measured with this technique vary from 10 to 250 μm (0.5 to 10 mils), resulting in an error ranging from 0.4 to 7.6 μm . (0.015 to 0.30 mils). We expect to reduce the total error by more extensive testing on electrodeposited chromium to evaluate v_c . The main problem lies in the non-uniform composition of the film, as well as the presence of inclusions and other kind of defects.

The temperature coefficient of sound velocity, β , was assumed the same for steel and chromium. The error in the estimated thickness due to this approximation can be expressed by the product of the uncertainty in the estimation of β ($\Delta\beta$), and in the difference in temperature between the standards and the tube. Even if this difference is 100 °C and $\Delta\beta$ is assumed to be as high as 10^{-4} , the error is less than 1%, thus it can be neglected in view of the larger error introduced by the uncertainty given by γ .

A comparison between thickness measurements performed during plating by our system and values

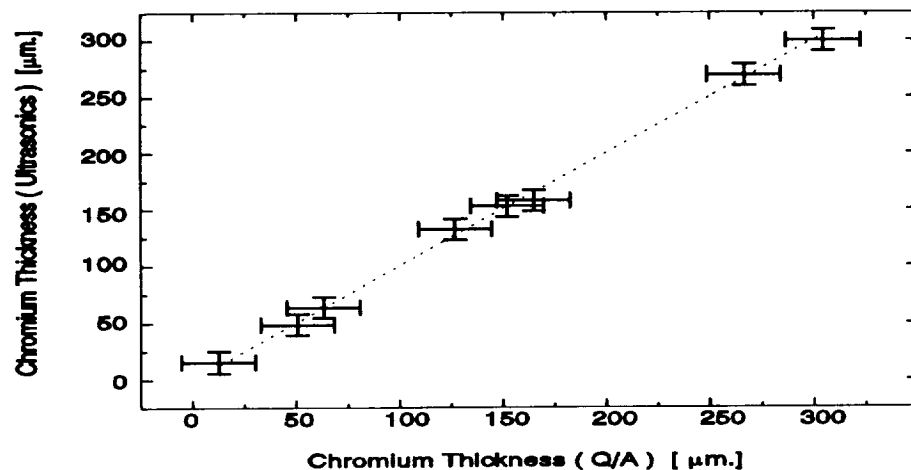


Figure 5 Error analysis.

later obtained by Quality Assurance (Q/A) personnel, was performed for an extensive number of shop plating runs. Results are shown in Fig. 5, where the thickness data obtained using our ultrasonic system is plotted as a function of the Q/A findings. The dotted line represents the one-to-one correspondence and the error bars represent the accuracy associated with each measurement. In the case of mechanical measurements performed by Q/A, this value is estimated to be $17.78 \mu\text{m}$ (0.0007 in.). It has to be pointed out that this is the accuracy required by process engineering. Good agreement between results can be found, even for higher values of chromium film thickness.

PRESENT AND FUTURE APPLICATIONS OF THE SYSTEM

The main advantage of our system over more conventional means utilized by Q/A personnel resides in the possibility of measuring the thickness of the chromium film as it is being plated as well as its greater accuracy. The system described has been implemented for many months in our vessel plating facility. In these cases, it has been possible to terminate the plating once the desired thickness of the chromium film had been reached, resulting in cost savings. Due to the strict tolerances of some plating processes, machining is required if too much chromium has been deposited. On the other hand, if the film is not thick enough, complete stripping of the chromium is required and the plating has to be performed a second time.

Furthermore, there were instances in which the measuring system was able to inform the plating operators of possible problems in the plating process. For example, in one case it was possible to make a quantitative determination of an interruption in deposition after a short interruption in current. Also the difference in plating rate before and after the interruption was measured, with the final thickness result again validated by Q/A personnel.

A problem that might be encountered in the plating of long tubes is straightness. Any curvature of the anode causes a different spacing between the anode and the inside diameter of the tube, resulting in different amounts of chromium deposited across the section. The system is able to detect the occurrence of this problem by comparing the thickness values obtained by the three transducer on the same ring.

All the measurements discussed here were obtained from test runs during the development of the ultrasonic thickness measuring system. Now, with the system fully developed and in use, plating is stopped or remedial action is taken as soon as any of the problems mentioned are detected. In the future, we envision a

computerized system, in which the information about the thickness of the chromium film is processed along with other information such as plating current or velocity of the liquid solution, to control and optimize the plating process. Further improvements of the system, from the point of view of process control and of increased accuracy is planned but not developed at this time.

SUMMARY AND CONCLUSIONS

Ultrasonics and computer technology have been integrated to obtain the means for "seeing" during plating. The thickness of a chromium film can be measured in real time with resolution and accuracy much better than that of the mechanical measurements which can be performed only after the plating process is finished. The advantages of this are many. Among those the possibility of intervention for remedial action in the case of problems. A detailed description of the system was given, starting from the theory of the measurement to the experimental details. The system is now being incorporated into process control so to stop plating as soon as the desired thickness is reached, or to monitor any unpredictable or irregular behavior of the process.

TABLE I. Results from Error Analysis		
Parameter	Standard Deviation	Error [μm]
Thickness of Standards [μm]	0.25	1.27
Velocity Ratio ($\gamma = v_c/v_s$)	0.03	6.42
Temperature Dependence β [$^{\circ}\text{C}^{-1}$]	10^{-5}	2.00
Time delay Meas. on Standards	.0001	6.22
Time Delay Meas. on Sample	.0001	1.39
Temperature Meas. [$^{\circ}\text{C}$]	1	0.75
	TOTAL [μm]	9.41

REFERENCES

1. J.C. Askew, U.S. Army Tech. Report ARCCB-TR-89021, (1989)
2. J.C. Askew, Proc. of this Conference
3. A. Abbate, J. Frankel and M. Doxbeck, Plating and Surface Finishing 80, 57, (1993)
4. J. Frankel, M. Doxbeck and A. Abbate, US Army Tech. Report ARCCB-TR-93001, (1993)
5. A. Abbate, M. Doxbeck, S.C. Schroeder and J. Frankel, Proc. Annual Review on QNDE, (1993)
6. E.P. Papadakis, Rev. Sci. Instr. 47, 806, (1976)
7. J. Frankel, W.J. Korman and G.C. Capsimalis, Proc. IEEE Ultr. Symp., 887, (1980)

S37-86

N94- 32457

2520

P. 8

**MICROWAVE SENSOR
FOR
ICE DETECTION**

BY

G.D. ARNDT*, A. CHU*

L.G. STOLARCZYK AND G. L. STOLARCZYK****

ABSTRACT

A microwave technique has been developed for detecting ice build-up on the wing surfaces of commercial airliners and highway bridges. A microstrip patch antenna serves as the sensor, with changes in the resonant frequency and impedance being dependent upon the overlying layers of ice, water and glycol mixtures. The antenna sensor is conformably mounted on the wing. The depth and dielectric constants of the layers are measured by comparing the complex resonant admittance with a calibrated standard. An initial breadboard unit has been built and tested. Additional development is now underway.

Another commercial application is in the robotics field of remote sensing of coal seam thickness.

1.0 INTRODUCTION

Ice build-up on the orbiter low temperature fuel tanks, airfoil surfaces and highway structures can create hazardous transportation conditions. Ice build-up on the orbiter's low temperature fuel tanks is a safety concern in NASA'S Space Shuttle Program. After filling insulated fuel tanks on the booster rocket, the count down time period can continue until ice build-up reaches 1/16 inch. During the insertion phase of flight, the ice layer could fragment and damage heat shield tile and windows of the shuttle. Presently, ice depth measurements are made manually by scratching away the ice layer and determining its thickness. Because of the large physical size of the fuel tanks, the number of measurements is limited.

- * NASA Johnson Space Flight Center
- ** Raton Technology Research, Inc.

Commercial airline disasters in Washington, DC, Denver, Colorado, New Foundland, New York and Europe have been caused by wing icing prior to the take-off role. Adherence of ice to the wing can reduce lift by as much as thirty-four percent and on some aircraft ice may dislodge and damage the jet engine. Formation of ice on airfoil surfaces caused 127 fatal commercial aircraft accidents and claimed 496 lives between 1977 and March 1992. All of these tragedies could have been prevented had there been a reliable device to determine whether deicing was needed. The Federal Aviation Administration (FAA) has projected that unless safety improvements are made, there could be eight additional air carrier icing accidents within the next ten years claiming 134 lives. In addition, the cost to improve safety operations during winter time could cost approximately \$181 million.

Type I anti-icing fluid is spread on the aircraft prior to take-off. In one airport alone, 700,000 gallons of highly toxic ethylene glycol are used during the winter season. The airport waste water treatment load is equivalent to a city of 500,000 people. Icing on highways is another national concern. Nearly 4 million miles of U.S. roadways and 500,000 bridges are subject to icing conditions at some time during the year. Almost one third of all U.S. traffic accidents occur on wet, ice or snow/slush covered roadway surfaces. The microwave ice detection sensor project will improve transportation safety by developing thin conformally mounted sensor technology. The sensor and associated microcomputer controlled electronics can determine the electrical parameters of the contaminant layer overlying the sensor. Fusion of data from surface temperature and piezo electric sensors enable complete characterization of the physical and electrical parameters of the contaminant layers.

Prior to the development of the microwave ice detection sensor, many technologies have been investigated and proven unreliable in measuring ice build-up under adverse weather conditions. One of the technical problems relates to the discrimination of ice, snow and water conditions on critical surfaces. Measurement of overlying layer depth is another problem. Protruding sensors may unreliably determine ice conditions on airfoil surfaces. Some sensor thermodynamic properties can also alter the ice forming characteristic near the monitoring point.

Theoretical and experimental studies of a resonant circular microstrip patch antenna sensor have shown that its measurable electrical properties are highly sensitive to the contaminant layer properties. The resonant frequency and terminal impedance of the patch antenna depend upon the layer depth, resistivity and dielectric constant of the overlying layer. Laboratory tests have shown that the sensor and associated electronics can discriminate between water, ice, snow/slush and antifreeze layers. Measurements can determine the thickness of the layer, as well as the ethylene-glycol mixture. This paper will describe the microstrip patch

antenna sensor as well as the theoretical and laboratory measurement results.

2.0 THEORETICAL BASIS OF ANTENNA SENSOR

The microstrip patch antenna which is used for ice layer measurements has its theoretical basis in a sensor suggested by Chang and Wait for measurement of uncut coal layer thickness. The Chang and Wait sensor consists of a resonant loop antenna positioned over a thin coal layer as illustrated in Figure 1.

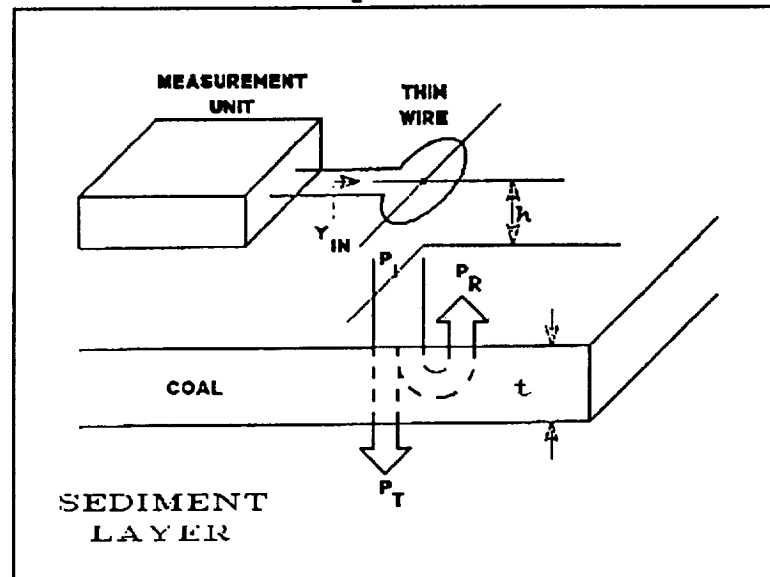


Figure 1 Uncut coal sensor.

The sensor measures the uncut coal layer thickness by measurement of the scattering of EM wave energy from the uncut and hidden surface. An antenna launches an incident EM wave (incident power [P_I]) that enters the uncut surface and travels a distance (t) to the sediment layer. A small part of the incident wave energy propagates (transmitted power [P_T]) into the sediment layer. The scattered wave (reflected power [P_R]) depends upon the thickness and the EM wave propagation constants of the uncut coal layer. Scattered wave energy returns to the antenna where it alters the admittance of the antenna.

Chang and Wait (1977) describe the theory of a resonant loop antenna operating over a coal layer covering the front of a thicker sediment layer. The conductivity (σ) and permittivity (ϵ) of the sediment layer is greater than that of coal. The coal and sediment layers form a two-layer half space with the coal layer of finite thickness and the second layer of infinite thickness. Chang (1973) developed formulations of the self admittance by applying a voltage of amplitude V_0 across the terminals of the loop antenna

illustrated in Figure 1. The perfectly conducting wire radius is assumed to be very small as compared with both the radius of the loop and the height (h) above the layered half space. The admittance of the loop at the feed point is given by:

$$Y = I/V_0 = G + iB \quad (1)$$

$$= I_0 + 2 \sum_{n=1}^{\infty} I_n \text{ Siemens} \quad (2)$$

where V_0 is the applied test voltage,

I is the current,

$$I_{n,1} = i 120\pi^2 (a_n^p + a_n^s) \quad (3)$$

a_n^p is the primary contribution of an isolated loop in the absence of the half space,
and, a_n^s is the contribution due to the half space.

The conductance (G) corresponds to the power radiated from the loop. The susceptance (B) of the loop is proportional to the amount of reactive energy stored in the vicinity of the antenna. King (1969) has developed a mathematical expression for a_n^p . For the case of a resonant loop, $|a_n^p|^{-1}$ is the largest term and current on the loop is basically cosinusoidal. The radiated power is greater at the resonant frequency than at a lower frequency. A resonant loop is expected to be more sensitive to thickness than a non-resonant loop.

To account for the influence of the layered half space, a_n^s needs to be included into the expression for current. Chang and Wait (1977) have used transcendental equations to evaluate the behavior of a_n^p and a_n^s for different heights of the coal layer and operating frequencies near resonance. As the uncut layer thickness increases, the conductance G exhibits a damped sinusoidal characteristic.

3.0 PRELIMINARY LABORATORY TESTS WITH PATCH SENSOR

Theoretical and experimental investigation of the resonant microstrip antenna sensor found that the percentage change in resonant frequency and conductance due to overlying ice and water layer depth could be measured with a practical instrument. The thin microstrip antenna sensor is illustrated in Figure 2.

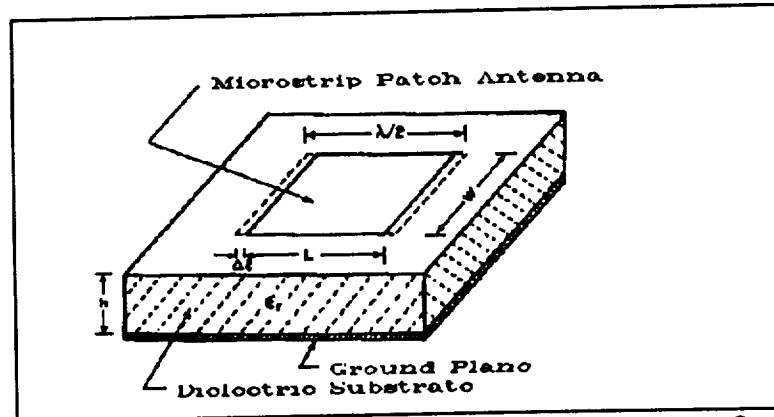


Figure 2 Resonant antenna structure for measuring ice thickness layers.

Computer codes were developed and used to determine the resonant frequency change of a microstrip antenna due to ice buildup. The theoretical results are illustrated in Figure 3.

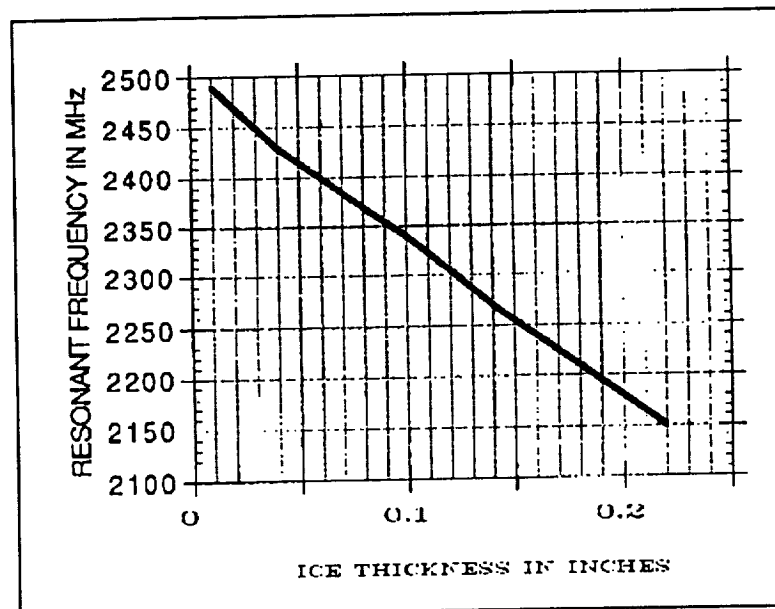


Figure 3 Theoretical resonant frequency versus ice thickness.

The resonant frequency changed by 140 MHz (5.6 percent) for a 0.1 inch change in ice layer depth. A 0.25 inch layer of ice caused the resonant frequency of the antenna to decrease by 14 percent. To investigate the ice and ice-water layering behavior in detail, a series of experimental tests were conducted in a temperature controlled chamber. In these tests, 0.1 inch depth increments of water were added to a tray in which the microstrip antenna sensor formed the bottom of the tray. The resonant frequency and conductance were independently measured after each incremental

change in water depth. The measurements were repeated after one hour when 0.1 inch water layer turned to ice. The test data is illustrated in Table A.

TABLE A
CIRCULAR MICROSTRIP ANTENNA
RESONANT FREQUENCY AND RESONANT CONDUCTOR
VERSUS ICE AND 0.1 WATER-ICE DEPTH

ICE DEPTH INCH	ICE		0.1 INCH WATER AND ICE	
	f_o (MHz)	G(mS)	f_o (MHz)	G(mS)
0.0	821.21	13.4	797.09	21.5
0.1	812.98	15.4	784.14	28
0.2	815.28	17.0	786.21	30.7
0.3	812.07	17.7	782.0	39.4
0.4	805.02	22.3	772.78	55.4

The measured data is illustrated in Figure 4.

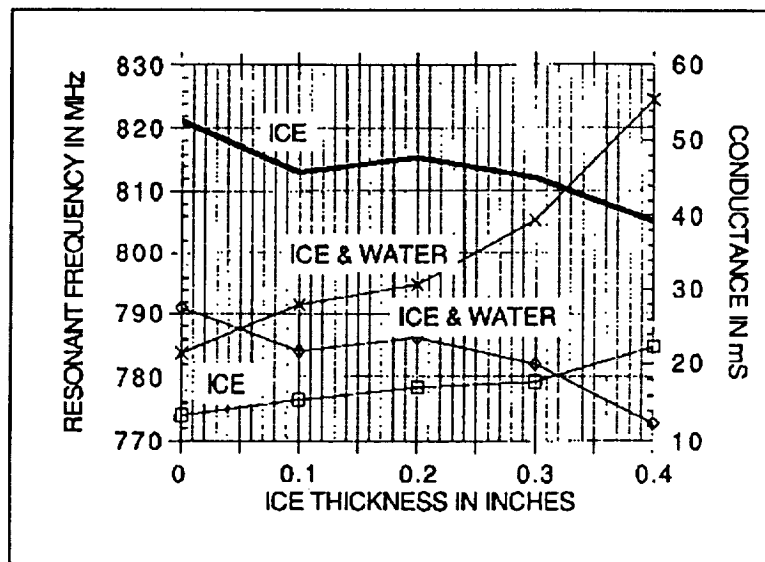


Figure 4 Measured resonant frequency and conductance vs ice thickness at 10° F.

The dark solid and square symbol curves indicate the change in resonant frequency and resonant conductance versus ice layer depth. The diamond and cross symbol curves represent the change in resonant frequency and conductance when 0.1 inch depth of water covers the antenna. The dark solid and diamond resonant frequency curves show that the frequency is significantly changed when a water layer is present.

The application of the sensor array on air transport is illustrated in Figure 5.

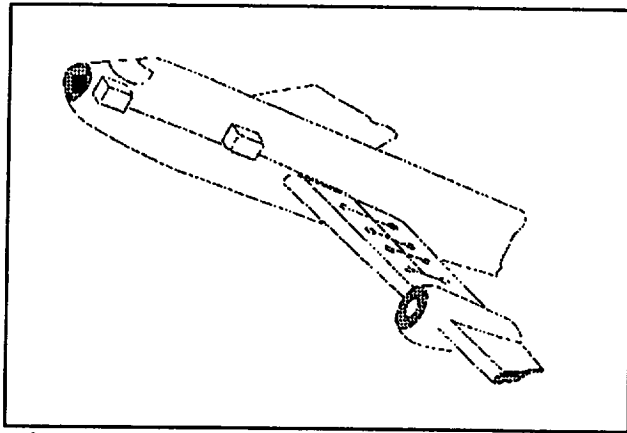


Figure 5: Perspective view of sensor array on airfoil surface

The critical surface of the wing extends outward from the fuselage to approximately 30 percent of the wing. It extends a distance of approximately 20 percent of the wing width. The thin conformal antenna array is multiplexed through a flat transmission line to the microcomputer controlled electronics in the fuselage of the aircraft. The electronics generates the flight deck display illustrating wing icing conditions. Since sensor measurements are continuous, the display can indicate time to freezing of the overlying layer.

SUMMARY

Theoretical and experimental research work has found that the contaminant layer properties can be measured with the sensor. The thin conformal sensor can be applied to the critical areas of the wing without altering its aerodynamic properties. Fusion of wing surface temperature measurements data with sensor data will enable estimation of time to freeze. Fusion of piezoelectric data will enable cross check of ethylene-glycol water mixer and freezing point. Adherence of the layer to the wing surface can also be detected with this type of sensor. The sensor fusion goal is to maximize reliability of ice detection for transportation safety.

REFERENCES

1. Stolarczyk, et al, U.S. Patent 4,753,484, "Method for Remote Control of Coal Shearer"; June 28, 1988.
2. Chang, D.C. and Wait, J.R., "An Analysis of a Resonant Loop as an Electromagnetic Sensor of Coal Seam Thickness", Proceedings 1977, URSI Conference on Remote Sensing, LaBelle, France.

3. Chang, David, "Characteristics of a Horizontal Circular Loop Antenna over a Multilayered Dissipative Half Space", IEEE Transactions on Antennas and Propagation, November 1973.
4. Gupta, K.C. and Benalla, A., Microstrip Antenna Design, Norwood, MA: Artech House, 1988.
5. Bahl, I.J. and Bhartia, P., Microstrip Antennas, Dedham, MA: Artech House, 1980.
6. King, R.W.P. (1969), Loop Antenna for Transmission and Reception, Antenna Theory, Pt. 1. Chap. II, pp 458-482, McGraw-Hill, NY.

A VERSATILE NONDESTRUCTIVE EVALUATION IMAGING WORKSTATION

E. James Chern and David W. Butler
Materials Branch / Code 313
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

Ultrasonic C-scan and eddy current imaging systems are of the pointwise type evaluation systems that rely on a mechanical scanner to physically maneuver a probe relative to the specimen point by point in order to acquire data and generate images. Since the ultrasonic C-scan and eddy current imaging systems are based on the same mechanical scanning mechanism, the two systems can be combined using the same PC platform with a common mechanical manipulation subsystem and integrated data acquisition software. Based on this concept, we have developed an IBM PC-based combined ultrasonic C-scan and eddy current imaging system. The system is modularized and provides capacity for future hardware and software expansions. Advantages associated with the combined system are: (1) eliminated duplication of the computer and mechanical hardware, (2) unified data acquisition, processing and storage software, (3) reduced setup time for repetitious ultrasonic and eddy current scans, and (4) improved system efficiency. This concept can be adopted to many engineering systems by integrating related PC-based instruments into one multipurpose workstation such as dispensing, machining, packaging, sorting, and other industrial applications.

INTRODUCTION

To maintain global technological competitiveness, manufacturers must use a concurrent engineering approach that includes performance, manufacturing, process monitoring, inspection, service and maintenance to design their products. Nondestructive evaluation (NDE), which is an essential test and measurement technology for the advanced concurrent engineering approach¹, offers many advantages over traditional quality assurance measures. Recent advances in computer and electronic technology have increased the processing speed and capabilities of personal computers (PCs), and enabled the development of application specific plug-in expansion boards. A wide range of stand-alone devices also have been converted to this type of circuit boards such as analog to digital (A/D) converters, digital multimeters (DMMs), interface buses, etc. The PCs thus can become various engineering workstations by using these plug-in boards, through interface buses for stand-alone digital instruments or through an A/D card for analog instruments.

Ultrasonic C-scan²⁻⁴ and eddy current imaging techniques⁵⁻⁷ are the two most widely utilized NDE imaging techniques that are routinely used for qualification, monitoring, and control of manufacturing processes. The ultrasonic C-scan and eddy current imaging systems are of pointwise type systems⁸ that have to rely on a mechanical scanner to physically maneuver a sensing probe relative to the specimen point by point in order to acquire data and generate images. Since these two imaging systems are based on the same scanning mechanism, the two systems can be combined using the same PC platform with a common mechanical manipulation subsystem and integrated data acquisition software.

In this paper, we describe the hardware/software requirements and the development of an IBM PC-based combined ultrasonic C-scan and eddy current (UT/EC) imaging workstation. This UT/EC pointwise imaging system includes a common mechanical control, data acquisition and image processing software. Measuring instruments are incorporated into the PC-platform by using an analog ultrasonic pulser/receiver and a digital eddy current impedance analyzer, a plug-in eddy current instrument, a digital multimeter, and an IEEE-488 interface card. Advantages associated with this integrated test and measurement system are: (1) eliminated duplication of the computer and mechanical hardware, (2) unified data acquisition, processing and storage software, (3) reduced setup time for repetitious ultrasonic and eddy current scans, (4) improved system efficiency, and (5) provided capacity for future hardware and software expansions.

HARDWARE REQUIREMENTS

The hardware of a typical automated NDE system has three major components: a system controller, a mechanical scanner, and appropriate instruments with associated sensors. A PC or a microprocessor based

instrument is normally used as the system controller. The system controller is responsible for controlling the instruments, commanding mechanical movements, acquiring signals, processing data and generating images. A mechanical scanner and associated hardware fixtures such as probe and specimen holders are used to relatively scan the sensor over the area of interest on the specimen. Dedicated instruments and electronic devices are used to excite a sensor such as an ultrasonic transducer or an eddy current probe, and measure desired parameters within the corresponding signals. A typical ultrasonic C-scan imaging system utilizes a pulser/receiver as the excitation and measurement instrument, whereas a typical eddy current system utilizes an impedance analyzer or an eddyscope. Various types of sensors such as focused and flat ultrasonic transducers, and absolute and differential eddy current coils with appropriate geometric configurations for specific applications are used as probes. The block diagram of the combined hardware structure is shown in Figure 1.

The developed UT/EC imaging system has many specific hardware components. A CompuAdd 325 PC (IBM-386 Compatible) is used as the system controller. A Delta Tau Data Systems' Programmable Multi-Axis Controller (PMAC) motion controller card, Compumotor Plus drivers/motors, and a Daedal X-Y linear table with incremental linear encoders make up the mechanical scanner. For the ultrasonic instrument subsystem, a 18"W x 30"L x 6"H Plexiglas immersion tank is mounted on the mechanical table. A Panametrics 5052UA pulser/receiver with gated peak detector and a Keithley 196 System DMM act as the drive and measuring instruments for the ultrasonic signals. For the eddy current instrument subsystem, a Hewlett-Packard 4194A Impedance/Gain-Phase Analyzer provides the signal drive and measurement capabilities for the eddy current signals. An SE Systems' SmartEddy 3.0 eddy current plug-in instrument is also implemented for other eddy current applications.

The system controller interfaces with the mechanical scanner, using the PMAC controller card and the SmartEddy 3.0 through the industry standard architecture (ISA) PC-bus. A National Instruments AT-GPIB controller card which also resides on the PC-bus, is used for digital instrument control. The ultrasonic amplitude data from the System DMM or the eddy current impedance data from the Impedance/Gain-Phase Analyzer is transferred to the system controller through the IEEE-488 general purpose interface bus (GPIB). A sketch of the system is shown in Figure 2.

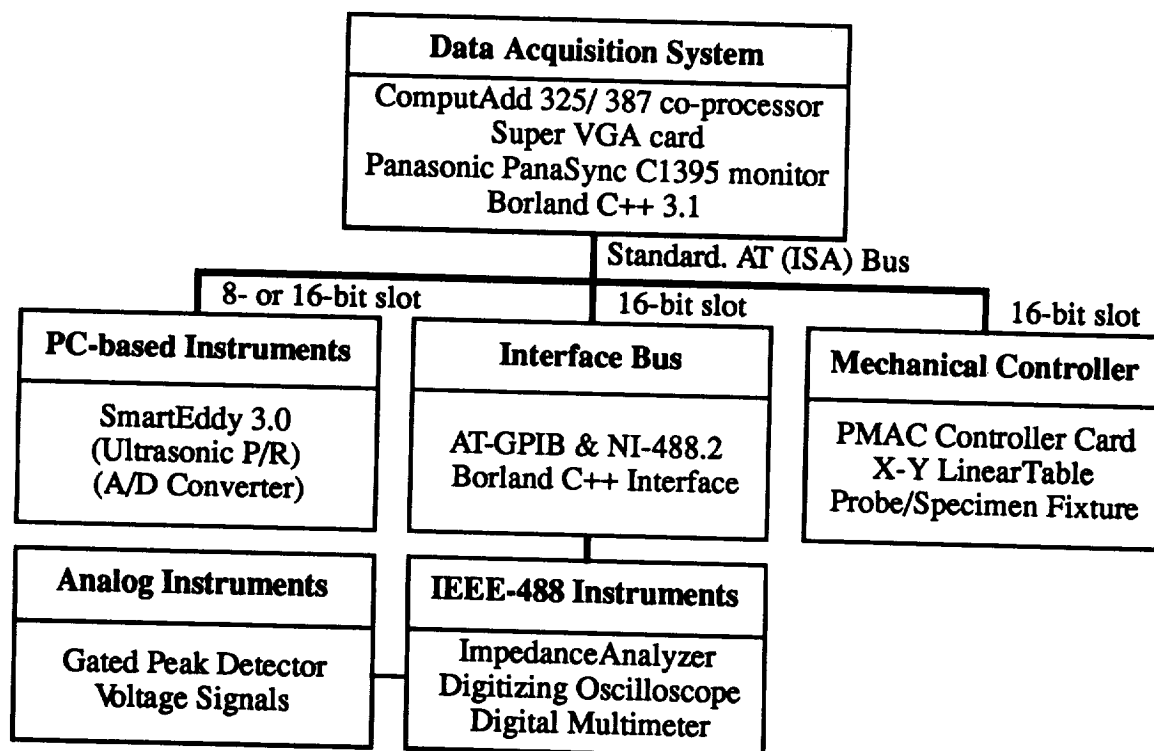


Figure 1. Block diagram of the UT/EC imaging workstation

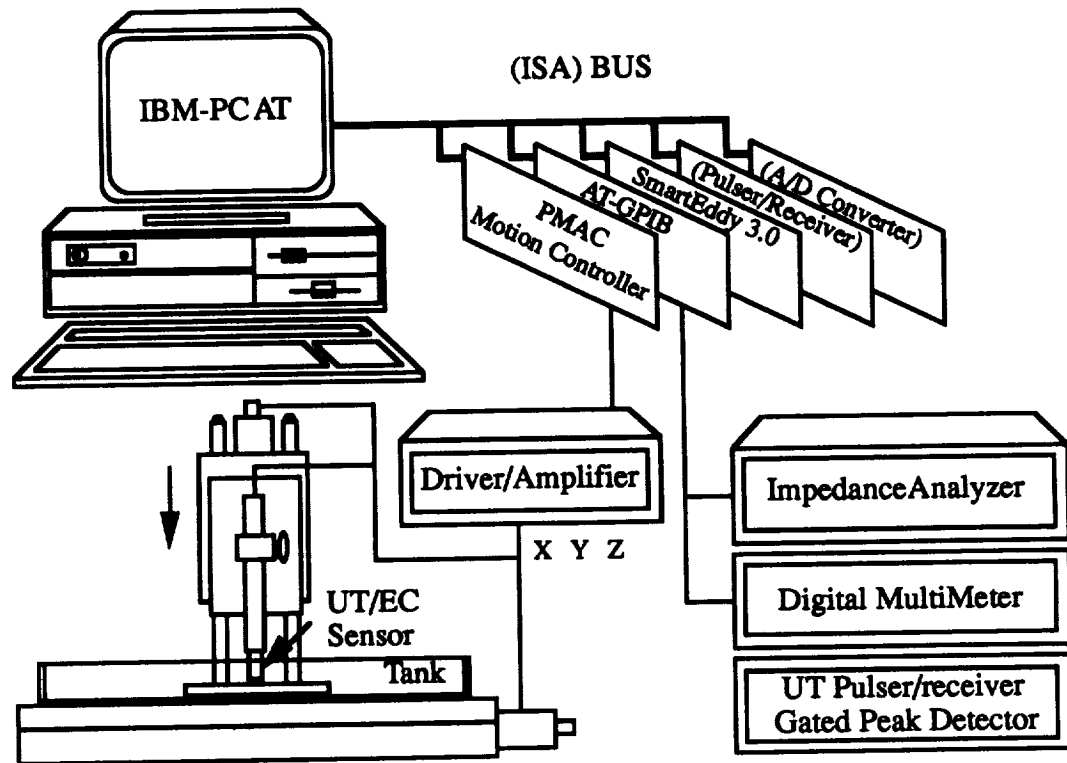


Figure 2. A sketch of the UT/EC imaging workstation showing essential components

SOFTWARE/FIRMWARE PROGRAMS

Software Development

The software programs of the UT/EC imaging system were developed with the Borland C++ 3.1 C compiler and other off-the-shelf development tools. The software module MAIN.C is the application program driver. Under this system driver program, software routines are structured into three groups as illustrated in Figure 3. The first group is for the GPIB hardware interface bus which is based on the National Instruments' NI-488.2 driver software, and user interface display which is based on the Ultra Windows user interface library package. The second group is the core of the data acquisition and management software. Modularized programs are devised for each engineering function of the inspection operation: initialization, inspection parameter inputs, instrument setups, mechanical control, data acquisition, image presentation, message displays, and archives. The third group is for the definition of data structures, function prototypes, and global variables as well as the color SVGA/VGA graphics drivers.

Firmware Development

The system controller of an ideal digital pointwise imaging system would command the scanner to scan at a desired speed, monitor its position with the encoder reference, and fetch measurements at the desired coordinates on the fly. A firmware approach which uses interrupts generated by the mechanical system, at the desired positions, to trigger and initiate the data acquisition routine is thus developed⁹ to satisfy this requirement. Interrupt handling routines were developed using the Borland C++ 3.1 programming package. The control program enables the PMAC to generate interrupts in the system controller as trigger signals to initiate data acquisition. The scan routine downloads the scan parameters to the PMAC memory along with the position of the first measurement. The remainder of the measurement positions are calculated "on the fly" during the scan, relative to the starting position.

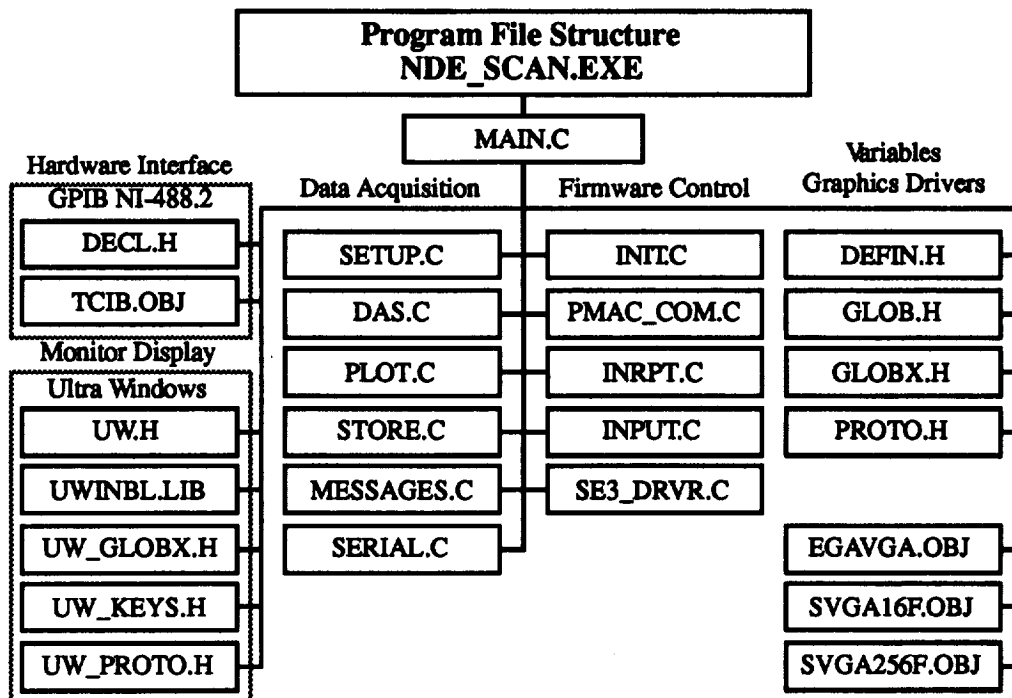


Figure 3. The file structure of the data acquisition software programs

While scanning, the PMAC microprocessor constantly compares the real-time probe position from the encoder feedback with the precalculated acquisition coordinates. An interrupt is generated by a Programmable Interrupt Controller (PIC) on the PMAC card and received by another PIC in the PC when the positional conditions are met. The PC PIC subsequently generates an interrupt to PC CPU. This interrupt is used by the CPU as the trigger signal to invoke the data acquisition routine and synchronize other events. The PMAC PIC continually generates interrupts until the scan routine is completed. Since the encoders are independent of the mechanical drives, the interrupts are generated precisely at the desired coordinates. The interrupt structure between host and peripheral PIC is shown in Figure 4.

System Operation

Because this UT/EC imaging system is a turn-key menu driven evaluation system, it is relatively easy to operate. Operators still need to be trained and certified according to specific procedures and requirements in order to conduct routine inspections. In the operation of this system, an operator is also required to have some basic understanding of IBM-PC or its compatibles and the disk operating system (DOS). Familiarization with DOS and the PC will greatly facilitate system operation and other housekeeping tasks. After setting up the instruments, one can follow the procedure and proceed with the evaluation. Upon execution of the executable NDE_SCAN.EXE file, the program provides the operator with a master menu for desired operations. The software program is interactive and will guide the operator through various functional levels and prompt for all the necessary parameters. The first level of operations include *Data Acquisition*, *Plot Data from File*, *File Management*, *Operating System*, and *Exit Program*. A sample of the menu screen is shown in Figure 5.

The second level of operation for *Data Acquisition* includes various ultrasonic and eddy current configurations such as raster scan, rotational scan, single-gate, dual-gate, etc. The third level of the data acquisition is the setup of scan parameters associated with the selected scan scheme and the execution of the scan. *Plot Data File* includes the selection and display of the stored data files. *File Management* includes file manipulation, data processing, and file conversion. *Operating System* provides a window of possibility to perform necessary DOS system operations within the NDE_SCAN program.

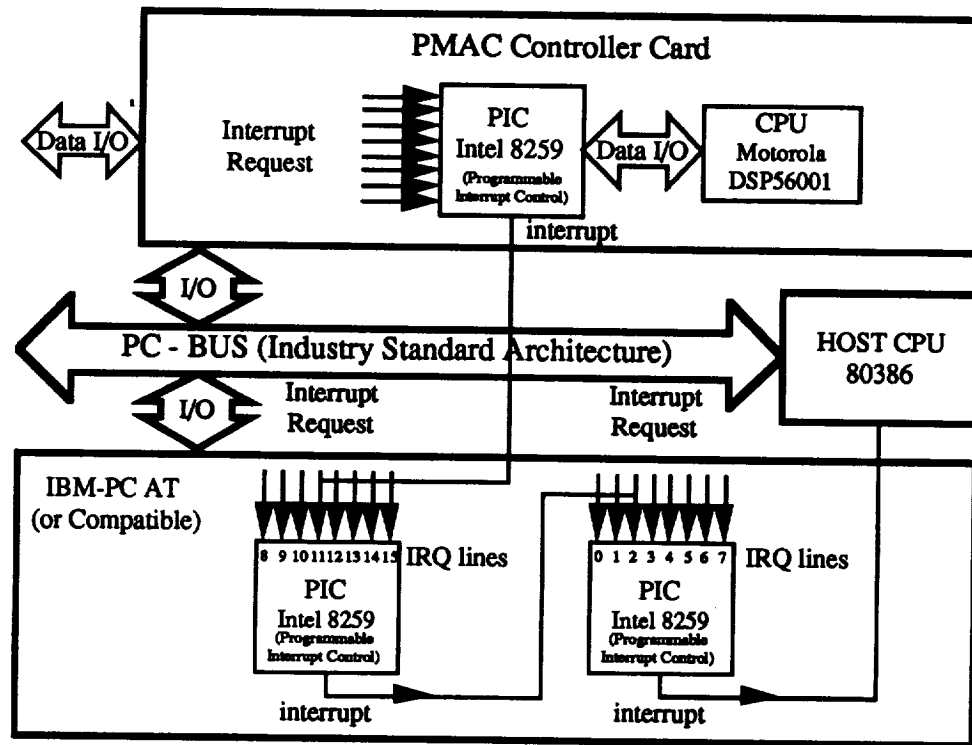


Figure 4. Interrupt structure of the host PC and PMAC PICs

ULTRASONIC/EDDY CURRENT IMAGING SYSTEM	
MAIN MENU	
Press Function Key or Letter to Make Selection	
<F1>	<u>D</u> ata Acquisition
<F2>	<u>P</u> lot Data From File
<F3>	<u>F</u> ile Management
<F4>	<u>O</u> perating System
<ESC>	<u>E</u> xit Program

Figure 5. The display of the MAIN MENU screen

CONCLUSION

In summary, we have successfully integrated a prototype versatile NDE imaging workstation based on off-the-shelf instruments and components. Since NDE instrumentation is a multidisciplinary field, the integration effort required certain specialized engineering expertise. The system is versatile and provides the flexibility to adapt specific inspection requirements and expansion needs. The benefits certainly justify the means.

There are many advantages associated with the combined system: (1) eliminated duplication of the computer and mechanical hardware, (2) unified data acquisition, processing and storage software, (3) reduced setup time for repetitious ultrasonic and eddy current scans, (4) versatility and flexibility, and (5) improved system efficiency. However, there is also a minor drawback. Currently, the system can only perform either ultrasonic C-scan or eddy current imaging at a given time. As it is required to calibrate the instruments for each inspection procedure. The throughput may be compromised if it needs to change back and forth between ultrasonic and eddy current scans.

The hardware and software of this NDE imaging workstation are readily available for technology transfer and marketing. Besides, IBM-PCs and their compatibles are gaining in popularity as system controllers and host computers for mechanical control, instrument control, and signal processing boards. IBM-PC based manufacturing, test and measuring systems are thus routinely being developed, introduced and implemented in various industries. This concept can be adopted to many engineering systems by integrating related PC-based instruments into one multipurpose workstation such as dispensing, machining, packaging, sorting, and many other industrial applications.

REFERENCES

1. E. J. Chern, "Concept of Nondestructive Evaluation," *Materials Evaluation*, Vol. 49, No. 9, pp. 1228-1235 (1991).
2. H. Berger, "A Survey of Ultrasonic Image Detection Methods," *Acoustical Holography*, Vol. 1, Plenum Press, New York, pp. 27-48 (1969).
3. G. S. Kino, "Acoustic Imaging for NDT," *Proceeding IEEE*, Vol. 67, pp. 510-523 (1979).
4. B. R. Tittmann, "Imaging in NDE," *Acoustical Imaging*, Vol. 9, Plenum Press, New York, pp. 315-340, (1980).
5. D. C. Copley, "Eddy Current Imaging for Defect Characterization," *Review of Progress in Quantitative NDE*, Vol. 3B, Plenum Press, New York, pp. 1527-1540 (1984).
6. W. G. Clark and B. J. Taszarek, "Stress Mapping with Eddy Currents," *Materials Evaluation*, Vol. 42, pp. 1272-1275 (1984).
7. E. J. Chern and A. L. Thompson, "Eddy Current Imaging for Material Surface Mapping," *Review of Progress in Quantitative NDE*, Vol. 6A, Plenum Press, New York, pp. 527-534 (1987).
8. E. J. Chern, "An Advanced Approach for Pointwise NDE Imaging," *Review of Progress in Quantitative NDE*, Vol. 12A, Plenum Press, New York, pp. 875-879, (1993).
9. E. J. Chern and D. W. Butler, "Firmware Development Improves System Efficiency," *Technology 2002*, The Third National Technology Transfer Conference & Exposition, NASA Conference Publication 3189, Vol. 1, pp. 425-434, (1992).
10. E. J. Chern, "An Integrated Ultrasonic and Eddy Current Imaging System," *Proceedings of the 6th Symposium on Nondestructive Characterization of Materials*, June 1993.

A NEW HIGH-SPEED IR CAMERA SYSTEM

**Jeffrey W. Travis
Peter K. Shu
Murzy D. Jhabvala
Michael S. Kasten
Solid State Device Development Branch
NASA/Goddard Space Flight Center
Code 718
Greenbelt, MD 20771**

**Samuel H. Moseley
Infrared Astrophysics Branch
NASA/Goddard Space Flight Center
Code 685
Greenbelt, MD 20771**

**Sean C. Casey
National Research Council
NASA/Goddard Space Flight Center
Code 685
Greenbelt, MD 20771**

**Lawrence K. McGovern
GSFC Space Academy
NASA/Goddard Space Flight Center
Code 685
Greenbelt, MD 20771**

**Philip J. Luers
Flight Data Systems Branch
NASA/Goddard Space Flight Center
Code 735
Greenbelt, MD 20771**

**Philip W. Dabney
Sensor Development and Characterization Branch
NASA/Goddard Space Flight Center
Code 925
Greenbelt, MD 20771**

**Ravi C. Kaipa
Hughes-STX
NASA/Goddard Space Flight Center
Code 718
Greenbelt, MD 20771**

**Nga Thuong Cao
NSI Mantech
NASA/Goddard Space Flight Center
Code 718
Greenbelt, MD 20771**

ABSTRACT

A multi-organizational team at the Goddard Space Flight Center is developing a new far infrared (FIR) camera system which furthers the state of the art for this type of instrument by incorporating recent advances in several technological disciplines. All aspects of the camera system are optimized for operation at the high data rates required for astronomical observations in the far infrared. The instrument is built around a Blocked Impurity Band (BIB) detector array which exhibits responsivity over a broad wavelength band and which is capable of operating at 1000 frames/sec, and consists of a focal plane dewar, a compact camera head electronics package, and a Digital Signal Processor (DSP)-based data system residing in a standard 486 personal computer. In this paper we discuss the overall system architecture, the focal plane dewar, and advanced features and design considerations for the electronics. This system, or one derived from it, may prove useful for many commercial and/or industrial infrared imaging or spectroscopic applications, including thermal machine vision for robotic manufacturing, photographic observation of short-duration thermal events such as combustion or chemical reactions, and high-resolution surveillance imaging.

INTRODUCTION

For astronomical imaging applications, large two-dimensional detector arrays can cover proportionally more sky in less time than single elements or linear arrays. Using a small telescope with a wide field of view and a suitably designed camera, astronomers can use such arrays to study the structure of star-forming regions in our own galaxy or distant galactic clusters. At visual and near-infrared wavelengths, the technology of silicon charge-coupled device (CCD) imagers is well understood, highly developed, and widely used, from astronomical instruments at the world's leading observatories to the hand-held video recorders found in many homes; arrays have been produced in formats as large as 4096x4096 pixels. Only recently, however, have two-dimensional arrays been developed for observations at far-infrared wavelengths. For example, the Rockwell Science Center has developed hybrid focal plane arrays (HFPAs) based on Blocked Impurity Band (BIB) technology. These HFPAs exhibit excellent responsivity over a broad wavelength band (see Fig. 1) and are capable of operating at 1000 frames/sec. [1, 2]

Ground-based and airborne observations at far-infrared wavelengths differ in two important aspects from comparable observations at visual wavelengths: i) the level of random background radiation from the telescope, as seen at the focal plane, is considerably higher than the level of radiation from distant astronomical targets; and ii) the relative level of atmospheric transmission can vary on the time scale of hours due to variations in line-of-sight water vapor. We have designed a high-speed camera system optimized for operation within these two constraints. To accommodate the high radiative background, we exploit the high-speed readout capability of the BIB array. A compact camera head electronics package continuously reads out and acquires data from the array at top speed, preventing it from saturating, while a digital signal processor (DSP) - based data system performs co-addition of frames in real time to improve the contrast. To monitor changes in atmospheric transmission, we have incorporated in our optical design a simple diffraction grating. This allows the instrument to operate as a spectrometer, and can be inserted into or removed from the optical path by means of a mechanical actuator.

Infrared detectors have been produced from a variety of materials and in a variety of formats to sense radiation from the near infrared ($1.0\mu\text{m}$) to the far infrared (longer than $20\mu\text{m}$). The choice of detector technology is driven by the specific requirements for the imaging task that needs to be accomplished. Although our instrument is designed around the BIB array, all of the electronics in the system have sufficient flexibility and modularity of architecture to accommodate other detector technologies for other applications.

SYSTEM OVERVIEW

A block diagram of the camera system is shown in Fig. 2. The BIB detector array and system optics are housed in a bench-top vacuum dewar. Infrared radiation from the telescope enters the dewar through a window in the outer case and is filtered by optical elements prior to focusing. The array is driven by a compact electronics package consisting of 6 boards: a timing generator, a clock driver/bias board, and 4 correlated double sampler (CDS) boards. This complement of boards provides all clocks and DC bias voltages required to operate the BIB array and digitizes its

analog output signals. Mounted on the outside of the dewar is a compact four-channel preamplifier module -- one channel for each output of the array. The preamplifier outputs are capable of driving 50 Ω coaxial cable with very fast settling times, allowing the rest of the camera head electronics to be located several feet from the dewar. Digitized data from the camera head electronics is sent to the data system over four optical fibers, allowing the data system to be located arbitrarily far away and isolating the camera head from electrical noise generated by the data system.

The data system consists of four custom DSP circuit boards residing in the backplane of an IBM-compatible personal computer (PC). Each DSP board processes digitized detector data from one output port of the BIB array. The data is accumulated in local memory on the DSP board and is collected, post-processed, displayed, and stored by the host central processing unit (CPU) following a complete observation.

FOCAL PLANE DEWAR

The camera/spectrometer optics and the BIB array are housed in a LHe reservoir dewar produced by Infrared Laboratories, lowering their temperature to approximately 4°K to reduce background radiation produced by the optics and thermally-generated signal produced by the BIB array. The cold plate of the reservoir is 8 inches in diameter. The optical path in the dewar is a folded, two-level design which attaches to the cold plate as a module and can be easily removed for alignment and test; we illustrate the design in Figure 3. The module is built up on a deck, also 8 inches in diameter, which divides the optical path into upper and lower sections for re-imaging and spectroscopy, respectively. The deck is elevated above the cold plate by means of precision machined, thermally conductive standoffs. The BIB array looks into the exit pupil of the optics module. Both sections of the optical path use pair-wise off-axis parabolic mirror segments. The segments are matched in a complementary fashion so as to cancel the aberrations normally incurred by a single off-axis mirror. The mirrors are fabricated from copper for high thermal conductivity, diamond-turned for an accurate figure, and gold-plated for high reflectivity.

The re-imaging portion of the optics module combines two off-axis paraboloids with differing focal lengths for a resultant focal plane magnification of 1.5. A single rectangular flat mirror folds the beam to accommodate the 8" diameter of the deck. The imaging section also includes a pupil stop and an adjustable filter slide. The pupil stop restricts the camera's field of view to the secondary and primary mirrors. Band-pass optical filters mounted on the filter slide restrict the spectral pass band of the camera for both imaging and spectroscopic observations. The parabolic mirror segments and a second beam-folding mirror form a second focal plane image at the deck's midplane. A bistable slide actuator positions a long-slit field stop in an opening at this location during spectroscopic observations.

The spectroscopic portion of the optics module contains two identical off-axis parabolic mirror segments and a second bistable actuator for a two-sided mirror/grating mount. The mirror/grating mount actuator is operated in conjunction with the field stop actuator. In the imaging configuration, the mirror re-images the midplane field stop onto the infrared active portion of the BIB array. In the spectroscopic configuration, an echelette grating is used to disperse light along the dimension of the restricted midplane field stop. Spot diagrams computed for the system suggest that the distortion of point sources at the outer edges of the focal plane's field of view are expected to lie within the area of a single pixel of the array. Our initial grating selection provides a dispersion of 20-40 μ m with a spectral resolution of $\lambda/\Delta\lambda \approx 30$.

The BIB array is located beneath the exit pupil of the optics module. Thermal coupling to the cold plate is made via a "cold finger", onto which the array package is gently clamped, while electrical connections are made by means of a printed circuit board with a socket which fits the array package. Individual coaxial cables are soldered directly to the printed circuit board to bring clock and bias signals to the board from the dewar electrical feedthrough and to return analog outputs to the preamplifiers. Simple transistor follower circuits on the printed circuit board protect the array outputs against external short circuits and reduce the output impedance of the array, allowing the transmission of high-speed analog signals over short lengths of coaxial cable to get to the external preamplifiers.

CAMERA HEAD ELECTRONICS

In the main camera head electronics package, the timing generator produces all of the digital signals required to operate the array and control the operation of the CDS boards. The circuit is based around the Am29CPL154 field programmable controller (FPC), produced by Advanced Micro Devices. This approach to timing generation is compact, flexible, and reprogrammable, and has a long heritage at GSFC. [3] The timing generator circuitry is completely isolated from all other camera head circuitry via optocouplers and an isolated power supply. Four clock signals are required to operate the BIB array; ten others are required to control the operation of the CDS boards. Four input lines are provided for interfacing to the data system or other external components.

The clock driver/bias board accepts logic level (0 to +5V) timing signals from the timing generator and shifts their voltage levels to meet the requirements of the BIB array multiplexer (+3 to +7V). It also generates the DC bias voltages required by the array. These voltages are referenced to a "virtual ground" which is at 0V with respect to the ground of the clock driver circuitry, but which is isolated from it. Both grounds are carried to the array fanout board in the dewar and kept separate, so that sensitive analog signals originating in circuits biased by the DC voltages are not corrupted by transient voltage noise on the clock lines.

The circuit design of the preamplifiers, and the way in which they are packaged and connected to the array, are critical to the preservation of the integrity of the analog output signals; the settling time (to 12-bit accuracy) of the analog signals, as measured at the input pins of the analog to digital converters (ADCs) must be in the 20-40ns range in order to fully realize the capability of the array. Low capacitance coaxial cables inside the dewar bring the buffered analog signals to the dewar electrical feedthrough, where they are mated with controlled-impedance connectors to get to the outside. The close proximity of the preamplifier package to the feedthrough assures a minimum cable run. The preamplifier package is carefully shielded against both radiated and conducted interference.

The CDS boards incorporate several significant new technologies to achieve a high degree of functionality in a compact form factor. The boards are designed around the ADS-119 ADC, produced by Datal. This part is a complete sampling 12-bit converter which employs a two-pass subranging scheme to achieve a maximum sample rate of 10MHz. The CDS board has sockets for two ADCs, and is capable of operating both at the maximum data rate that the BIB array will allow; however, in order to provide a "clean" interface -- from preamplifier to CDS board to DSP board, thereby simplifying integration and test -- we have chosen instead to populate only one of the two sockets. All analog circuitry on the CDS board, including the ADCs, is isolated from the digital side via optocouplers and isolated power supplies. The analog circuit ground is connected to the "virtual ground" generated on the clock driver/bias board via the preamplifier return lines from the dewar.

The analog signals from the preamplifiers are sampled twice per pixel period -- once for the reference level and once for the video level -- hence the term "correlated double sampling". The two conversion results are written into separate registers. The reference is then subtracted from the video via addition in ones complement form. The resulting 12-bit-plus-sign number is written out in two 10-bit words, along with synchronization bits from the timing generator. All of the registers and subtraction logic are implemented in a single field programmable gate array (FPGA) IC produced by Xilinx; all control signals for the CDS board (convert commands, register writes, multiplexer select lines, etc.) come from the timing generator. The use of the FPGA allows the CDS board a large measure of flexibility in its application. The FPGA can be configured for operation with one or both ADCs and single or double sampling. In addition, the FPGA can provide digital data to a digital to analog converter (DAC), located on the analog side of the board, for control of the DC level of the analog signal.

The 10-bit data words from the FPGA are written to a "Transparent Asynchronous Xceiver Interface" (TAXI) transmitter IC, produced by Advanced Micro Devices. The TAXI is mated with a fiber optic transmitter (available from several vendors), which incorporates interface circuitry, the optical transmitter element, and an "ST"-style fiber optic connector in a 16-pin dual in-line package (DIP). Sockets for differential TTL drivers are also provided for application in systems that do not employ fiber optics.

DATA SYSTEM ELECTRONICS

To handle the high data rate of our application, we have designed a custom circuit board around the ADSP21020 floating-point digital signal processor, produced by Analog Devices. The ADSP21020 utilizes a pipelined Harvard architecture (two identical independent buses, nominally for Program and Data) to achieve an instruction cycle time of 40ns at 25MHz. Its design includes a full complement of specialized circuitry, including an on-chip instruction cache, to optimize the processor for DSP applications, and it performs 40-bit floating-point operations according to the standard published by the Institute of Electrical and Electronic Engineers (IEEE).

The program memory data (PMD) bus has over 1.5 megabytes of 12ns static random-access memory (SRAM), organized as 256K words by 48 bits, for the storage of programs and co-added (accumulated) detector data. The word width is necessary to accommodate the 48-bit instruction format of the ADSP21020. The amount of memory on the PMD bus is sufficient to contain not only the DSP operating program, but also a number of independent frame accumulator areas, depending on the detector array being used and the nature of the DSP software (for example, each output of the BIB array requires only 4K words per frame accumulator); in addition, the PMD memory is expandable by mounting a "daughter board" containing the extra memory onto headers provided for that purpose. The data memory data (DMD) bus has 320 kilobytes of high speed SRAM, organized as 64K words x 40 bits, so that any floating-point computation may take place in the DMD space without interfering with real time accumulation in the PMD space.

Each DSP board receives detector data from a CDS board over a fiber optic cable using an integrated fiber-optic receiver interfaced to a TAXI receiver. The TAXI receiver is configured to receive 10-bit words at an effective rate of 8 million words/second. When the TAXI transmitter on the CDS board is not transmitting meaningful data, it continually transmits a synchronization pattern so that the transmitter and receiver remain in lock even when not in use.

Data from the TAXI receiver are stored into a 16K word x 18-bit first-in/first-out (FIFO) memory, two words at a time. The two spare bits are ignored. The Frame Sync, Chopper, and Sign bits are stored into the FIFO along with their accompanying data word; these bits are used by the DSP software to assign incoming data to the proper frame accumulator area. Reading a word of data with the Frame Sync bit set causes the highest priority interrupt to the ADSP21020. Reading a word of data with a Chopper bit that is different from the previous pixel causes a next lower priority interrupt.

The DSP board communicates with the host CPU via a set of registers mapped into the CPU's input/output (I/O) space. These registers appear in the I/O map in a space referred to in most documentation as "prototype board" at addresses 300H - 377H. Jumpers on the DSP board configure exactly where in this space the DSP board will reside to allow the four boards to be mapped into the prototype space without conflict.

HOST CPU & SOFTWARE

The PC platform we have assembled for the data system is a rack-mounting unit with a passive Extended Industry Standard Architecture (EISA) backplane. The CPU, like the four DSPs, is on a circuit board which plugs into the backplane. The CPU card incorporates an Intel 80486DX/50 microprocessor and 16 megabytes of RAM. Other system features include: a 1 gigabyte hard disk; a 128 megabyte magneto-optical cartridge drive; 5.25", 1.2 megabyte and 3.5", 2.88 megabyte floppy disk drives, and a Super VGA video adapter.

The data system software is a custom application which provides user interface and controls real-time data processing on the DSP boards. The software handles the details of data acquisition and processing, the transfer of accumulated data from the DSP boards to the host CPU, the user interface, and display, storage, and post-processing of the data. The DSPs are running a separate program from that running on the host CPU (a conventional microprocessor) and require their own unique development environment and programming language. However, the interaction between the two pieces of software is transparent to the user -- the appearance is only that the PC platform is collecting the data, processing and storing it.

Upon initiation of the application, identical executable machine code is loaded into the PMD space of each DSP board and execution initiated. This code processes operating commands from the host CPU, controls the fiber optic interface, retrieves the detector data, co-adds the data in the frame accumulators in real time, and communicates the results to the host CPU via registers which are mapped into the CPU's I/O space. The host CPU software allows the user to specify parameters for the acquisition (i.e., number of frames to be averaged, number of on target and off target frames, trigger conditions etc.) to the DSPs, waits for acquisition to be completed, and transfers the acquired data into the main system memory for image display, storage, and post processing.

CONCLUSIONS: INDUSTRIAL/COMMERCIAL APPLICATIONS

Infrared radiation is essentially a thermal phenomenon -- the hotter a body is, the shorter the wavelength of radiation it emits. Temperature differentials which are invisible to the eye can be converted into visible images on a computer screen if the scene is imaged onto an IR detector array and then digitized. Hence, IR detectors, from the very short wavelength IR ($1.0\mu\text{m}$) to the far IR (longer than $20\mu\text{m}$) find application anywhere where it is desired to sense thermal structure in a scene. IR detectors are commonly used in medical electronics, weather satellites, industrial process monitoring, food analysis, air traffic collision avoidance, air pollution analysis, monitoring thermal pollution from industrial effluents, and agricultural and oceanic biological content analysis, as well as a wide range of military and space applications. Generally, low-cost, rugged systems can be operated at ambient temperatures, whereas high performance, high sensitivity detection systems such as the one described in the present paper must be cooled to anywhere from 4 to 273°K , the exact temperature depending on the specific detector material and performance required. Although originally designed as an astronomical instrument, our camera system could be reconfigured for a variety of commercial and/or industrial IR imaging or spectroscopic applications, including thermal machine vision for robotic manufacturing, photographic observation of short-duration thermal events such as combustion or chemical reactions, and high-resolution surveillance imaging.

The application for which our system is designed is a very demanding one, hence the cost of the individual components required to meet the demands of the application, as well as the overall system cost, is high. However, the same basic system architecture could be implemented in less costly versions for other applications. Each of the major components -- detector array, optics, camera head electronics, data system hardware, and data system software -- could be selected, modified slightly, or redesigned to accommodate the requirements of the intended application.

The wavelength range of the BIB array may not be suitable for observing high-temperature phenomena; another IR technology, such as a platinum silicide (PtSi) photodiode array (roughly $1\text{--}5\mu\text{m}$), for example, might be more desirable. If another detector array were chosen, it might be possible to cool it only to liquid nitrogen (LN_2) temperatures (about 77°K), which would reduce operating costs considerably, or even to use a thermoelectric cooler (TEC) if the operating temperature were higher. Clearly, our modular approach to the optical design allows a great deal of flexibility in this area. Depending on the application, cold optics might not even be required at all.

In the camera head electronics, we have already discussed the reprogrammability of the timing generator and the reconfigurability of the CDS boards. If the application does not require the high speed at which we are operating, the preamplifier module and CDS boards are two areas where cost could be reduced by using lower speed parts. Likewise, for the DSP boards, we are using the highest speed parts available in order to meet our system throughput requirements; significant cost savings could be realized by using slightly slower parts if the application would allow for it. The expandability of the DSP memory is intended to provide a simple upgrade path for use with larger format arrays. Lastly, our data system software is completely flexible with regard to detector array format and measurement scenario, limited only by the amount of memory on the DSP board and in the host system.

REFERENCES

1. Huffman, J. E., Crouse, A. G., Halleck, B. L., Downes, T. V., and Herter, T. L., "Si:Sb Blocked Impurity Band Detectors for IR Astronomy", preprint (1991).
2. Sieb, D. H., Hays, K. M., Lin, W. N., Heimbigner, G. L., and Stetson, S. B., "Performance of 128 x 128 Element Switched Mosfet/Blocked Impurity Band Detector Hybrid Arrays", preprint (1991).
3. Hostetter, M., McCloskey, J., and Reed, K., "Microcontrollers Generate Timing Signals for CCD Arrays", NASA Tech Briefs, June, 1992, pp. 41-42.

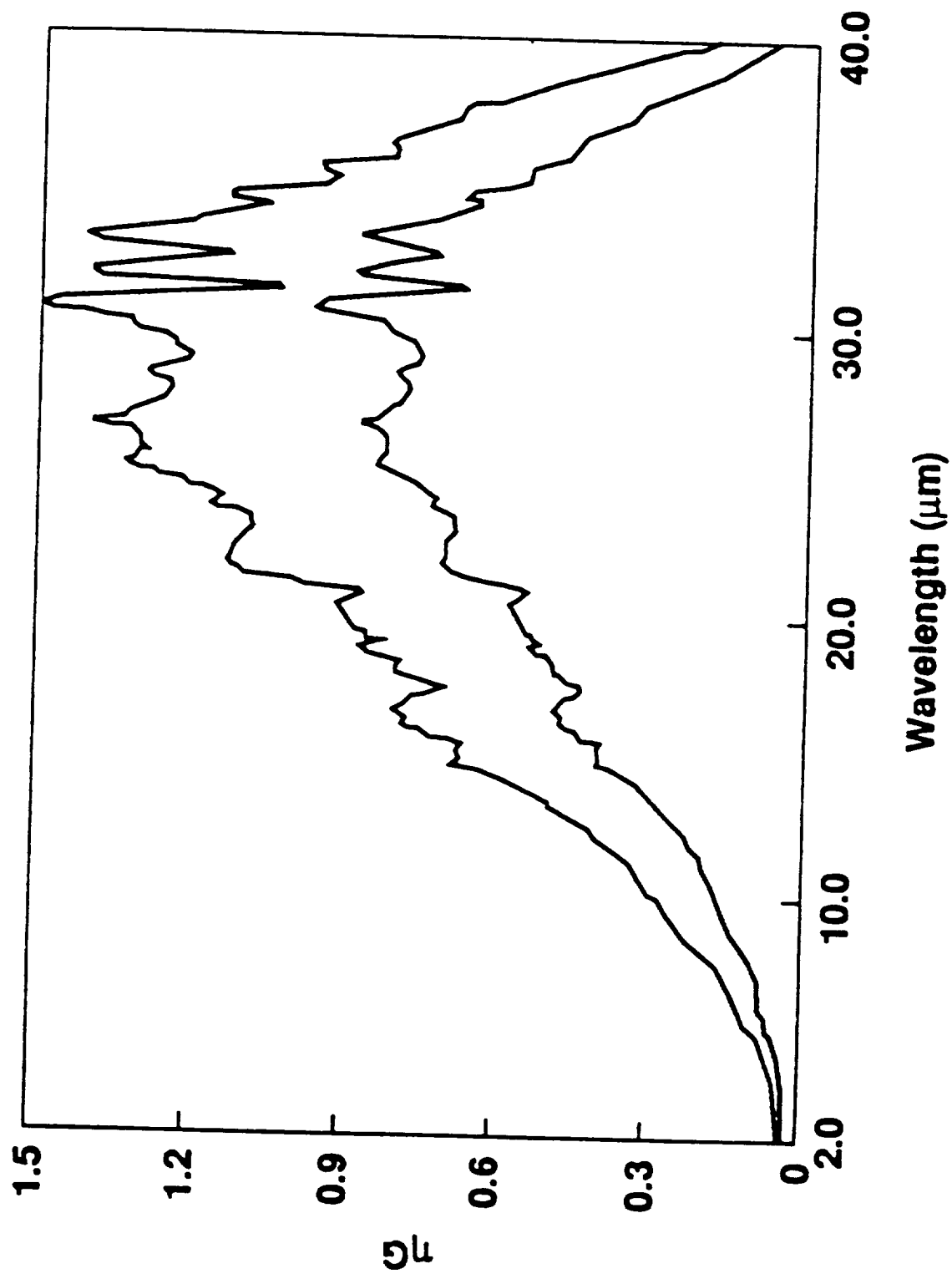


Figure 1. A figure taken from Huffman et al. (1991) on the spectral response of an earlier BIB detector design operating at two different bias voltages. The vertical axis notation, ηG (dimensionless), represents the relative responsivity of the device.

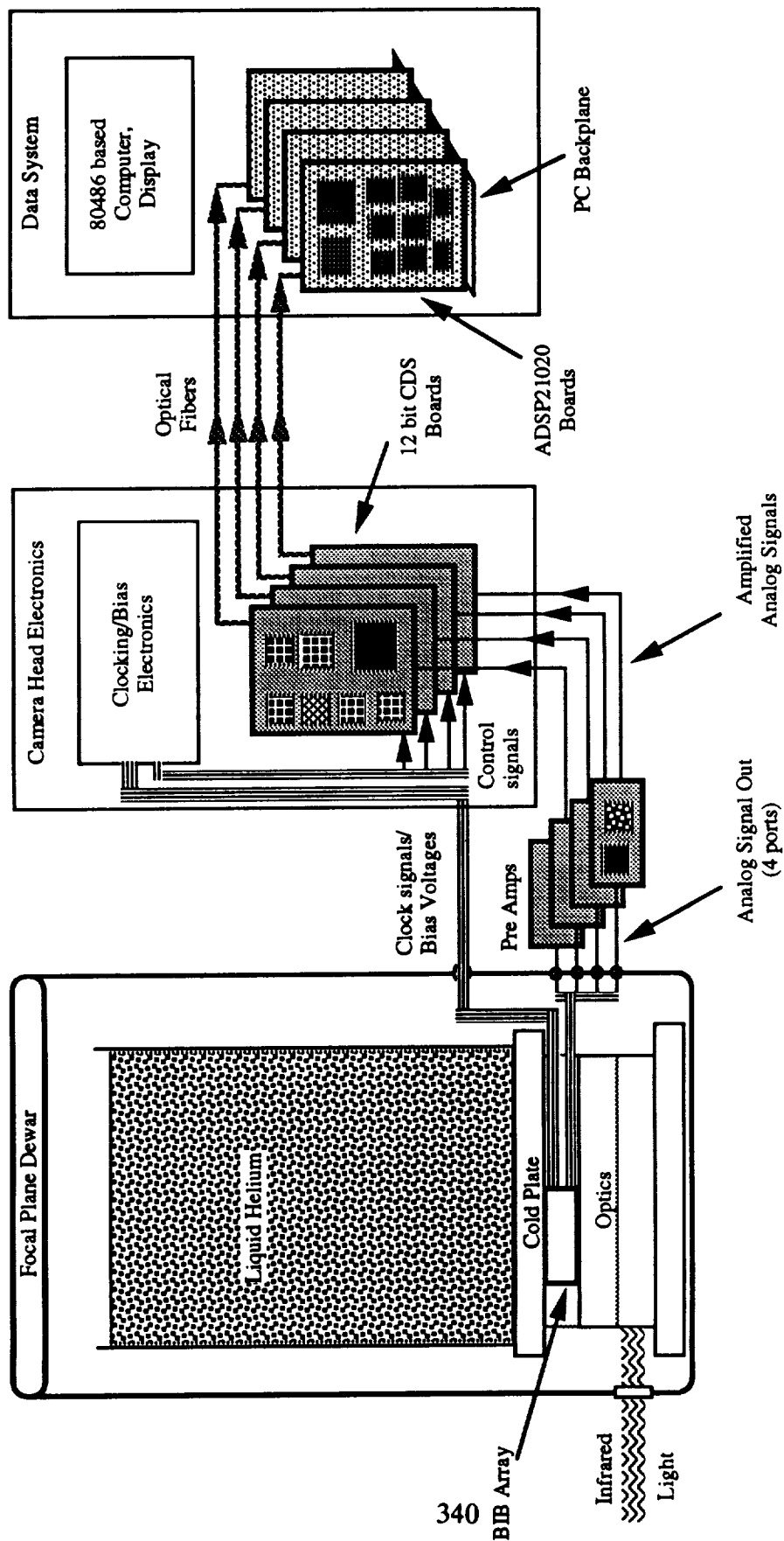


Figure 2. A block diagram of the BIB camera system. Note that the dewar is in the "operating" orientation; directional references in the text are with the dewar in the "assembly" orientation, i.e., top and bottom reversed.

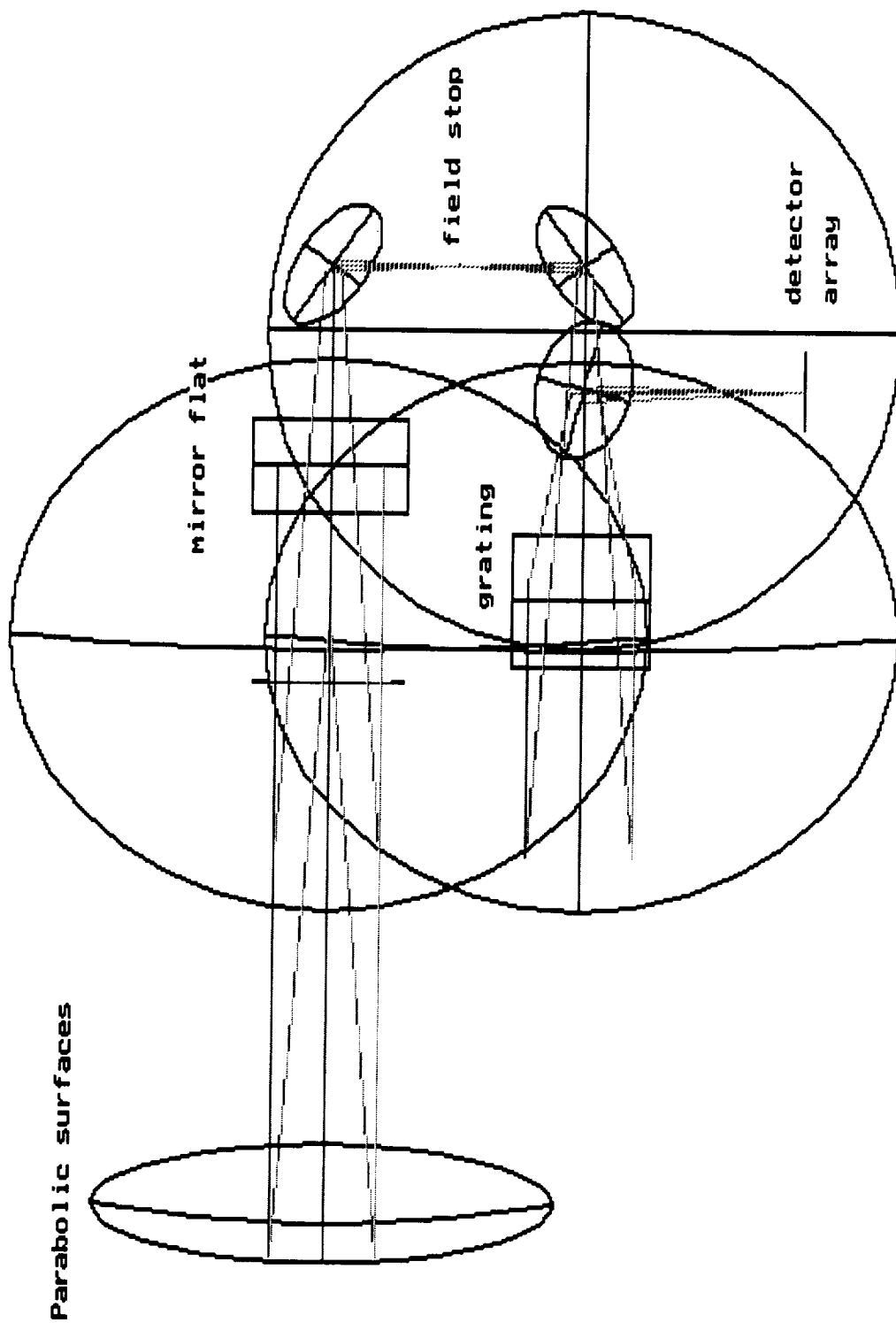


Figure 3. A schematic of the optical ray trace for our imaging/spectrometer camera design. The mirrors as depicted are the complete parabolic surfaces used by the ray tracing program BEAM4; only off-axis segments are used in the actual design.

540-33

2523

8-1

N94- 32460

UNIVERSAL SIGNAL CONDITIONING AMPLIFIER

Pedro J. Medelius

I-NET, Inc.

INI-10, Kennedy Space Center, FL 32899

Carl Hallberg

I-NET, Inc.

INI-3, Kennedy Space Center, FL 32899

Jim Cecil

NASA

DL-ESS-23, Kennedy Space Center, FL 32899

ABSTRACT

A state-of-the-art instrumentation amplifier capable of being used with most types of transducers has been developed at the Kennedy Space Center. This Universal Signal Conditioning Amplifier (USCA) can eliminate costly measurement setup time and troubleshooting, improve system reliability and provide more accurate data than conventional amplifiers. The USCA can configure itself for maximum resolution and accuracy based on information read from a RAM chip attached to each transducer. Excitation voltages or currents are also automatically configured. The amplifier uses both analog and digital state-of-the-art technology with analog-to-digital conversion performed in the early stages in order to minimize errors introduced by offset and gain drifts in the analog components. A dynamic temperature compensation scheme has been designed to achieve and maintain 12-bit accuracy of the amplifier from 0 to 70 °C. The digital signal processing section allows the implementation of digital filters up to 511th order. The amplifier can also perform real-time linearizations up to fourth order while processing data at a rate of 23.438 kS/s. Both digital and analog outputs are available from the amplifier.

DESCRIPTION

The Universal Signal Conditioning Amplifier is a self or remotely programmable amplifier which internally provides transducer excitation. The USCA was designed to improve the performance of the Permanent Measurement System (PMS) currently in use at the Kennedy Space Center. The USCA significantly reduces the time required to setup a new measurement, which currently takes several hours since amplifiers have to be physically matched to the transducers. Many transducers used in the PMS have outputs in the order of a few millivolts, and their amplifiers are sometimes several hundred feet away. Noise coupled into the cables can significantly deteriorate the performance of the measurement system. The USCA was designed to improve the signal-to-noise ratio by allowing the amplifier to be located close to the transducer when feasible. Each USCA is characterized over a 0 to 70 °C temperature range. By measuring the performance over the temperature range and by constantly monitoring the temperature, analog gain and offset drifts can be dynamically compensated by the digital signal processor stage.

With USCA, when a transducer is calibrated, a memory chip (TAG RAM) is attached to it. This memory chip contains information pertinent to the transducer, such as excitation levels, output range, linearization coefficients and others. Before a transducer is connected to the USCA, all the output voltages or currents are set to zero, and the input gain is set to unity. When the USCA is connected to a transducer, it sets itself up to adapt to the transducer by using the information stored in the Tag Ram. The default gains, excitation levels (voltage or current), filters, and output type (analog or digital) are set immediately upon connecting a transducer to the USCA. The default settings of the USCA can be changed remotely through the Self Aware Measurement System (SAMS) controller. The flexibility provided by the internal controller and the digital signal processing module permits the

use of non-standard transducers, such as pulse-type flow meters, sensors with frequency outputs, and A/C phase measurement transducers.

Self Aware Measurement System

A block diagram of a typical scenario for the use of the USCA is shown in Figure 1. A Self-Aware Measurement System (SAMS) controller is shown in the block diagram. The main purpose of this controller is to provide an interface between a host computer and the USCA. Even though the USCA sets itself up upon connection to a transducer, the SAMS controller can allow the selection of input and output gains. It also permits the selection of standard filters in each USCA, and allows the downloading of custom-designed digital filters. The SAMS allows for configuration control of the system by monitoring the number and types of transducers in use and by tracking calibration-due dates of the transducers. It maintains a date/time log of USCA and transducer connects and disconnects.

The SAMS controller can perform the configuration control even without a USCA present for transducers that do not require a signal conditioning amplifier, provided they are equipped with a Tag Ram.

Tag Ram

The Tag Ram consists of a memory chip, backed by a battery with a 10 year lifetime. This Tag Ram is loaded with data and sealed to the transducer after the transducer is calibrated. The Tag Ram includes information regarding the transducer type, required excitation level, output voltage range, calibration due date, linearization coefficients, ID number, and others. The Tag Ram is password-protected to prevent the calibration data of the transducer from being changed inadvertently. The communication between the USCA and the Tag Ram is done over a single pair of wires, and multiple Tag Rams can be connected to the single pair of wires. When long runs of cable are used to connect a USCA to a data acquisition system, each section of cable can be equipped with a Tag Ram. The SAMS controller then could perform the configuration control of the complete measurement, including the cabling.

USCA

The USCA combines state-of-the art analog and digital hardware to accomplish reliable and accurate signal conditioning. Figure 2 presents a block diagram which shows the main modules of USCA.

The input module consists mainly of a highly stable programmable gain amplifier. This module also includes transient and overvoltage protection circuits. The circuits have been designed to provide a DC to 10 kHz 0.02 dB passband with a 12 bit (1 part in 4096) accuracy. The programmable stage can select gains from 0.25 through 2000 V/V. The output of the amplifier stage is applied to a 16-bit analog-to-digital converter circuit which samples at 375 kS/s. A voltage reference with an stability better than 1 ppm/°C is used for the A/D converter. This is the most critical section of USCA, since the accuracy of the amplifier is mainly limited by the behavior of the analog section.

The excitation module provides a highly regulated voltage or current for the transducer. The excitation level is determined by the information read from the Tag Ram, and is controlled by a 16-bit digital-to-analog converter. This module allows the selection of an excitation voltage with a resolution of better than 500 μ V. The excitation stage has a current limiting circuit which restricts the maximum current to about 100 mA. The excitation can also be programmed to provide pulses with variable duty cycles and amplitudes.

The digital signal processing module performs the digital filtering and real-time linearization functions. This module consists of a Decimating Digital Filter (DDF) and a Digital Signal Processing (DSP) microprocessor. The DDF receives a data stream at a rate of 375 kS/s from the A/D converter. The decimator lowpass filters the input data and reduces the sampling rate to 23.438 kS/s or lower. The DDF also performs digital filtering up to 511th order using 20-bit coefficients. The available digital filters include lowpass, highpass, bandpass, and notch filters. The coefficients for the filters can reside inside USCA, or they can be downloaded through the SAMS controller. The filters designed for USCA allow for passbands with ripple lower than 0.02 dB. While running at 23.438 kS/s, the DSP processor can perform linearizations up to fourth order using the coefficients read from the Tag Ram. When running measurements which do not require such a fast sampling rate, larger order real-time linearizations can be performed.

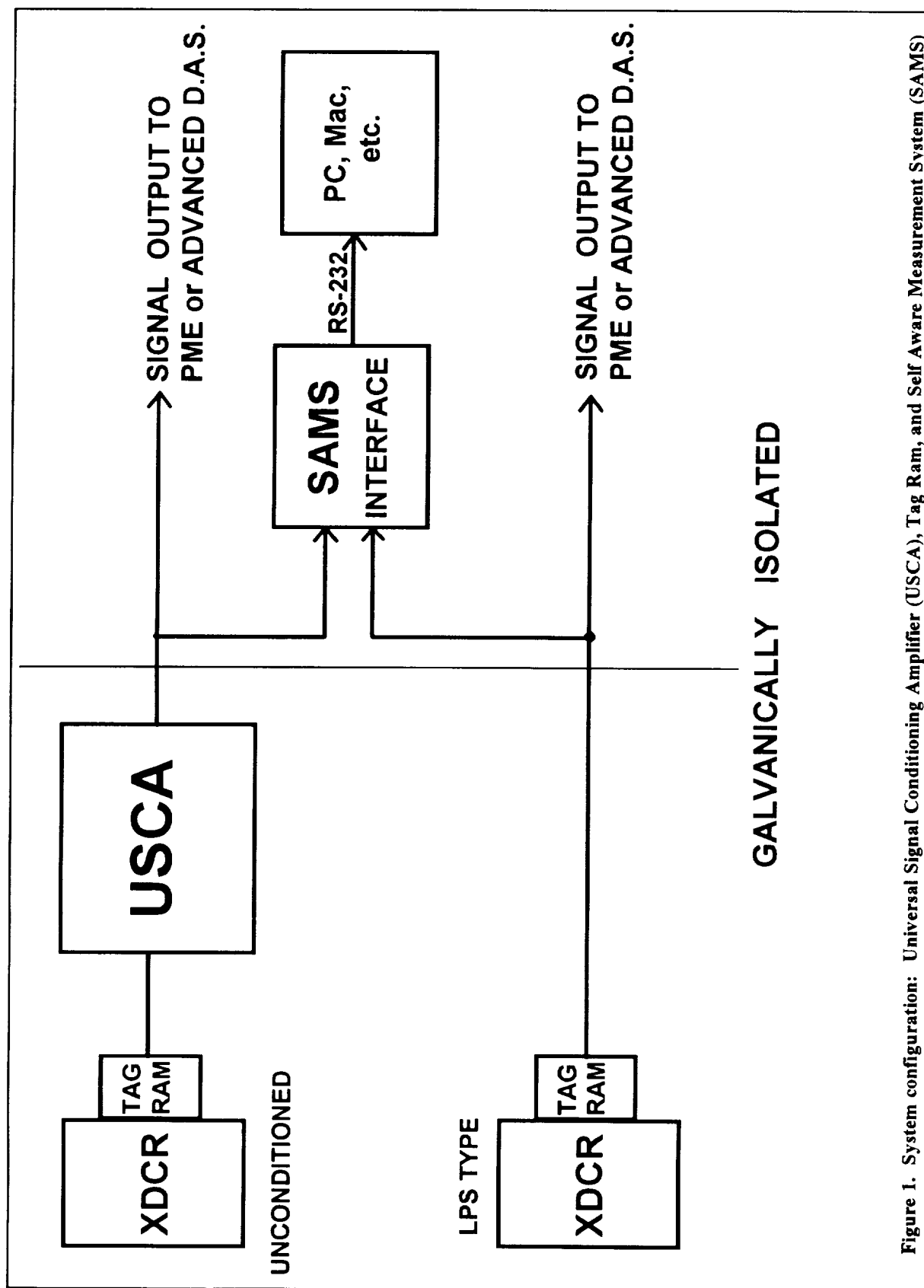
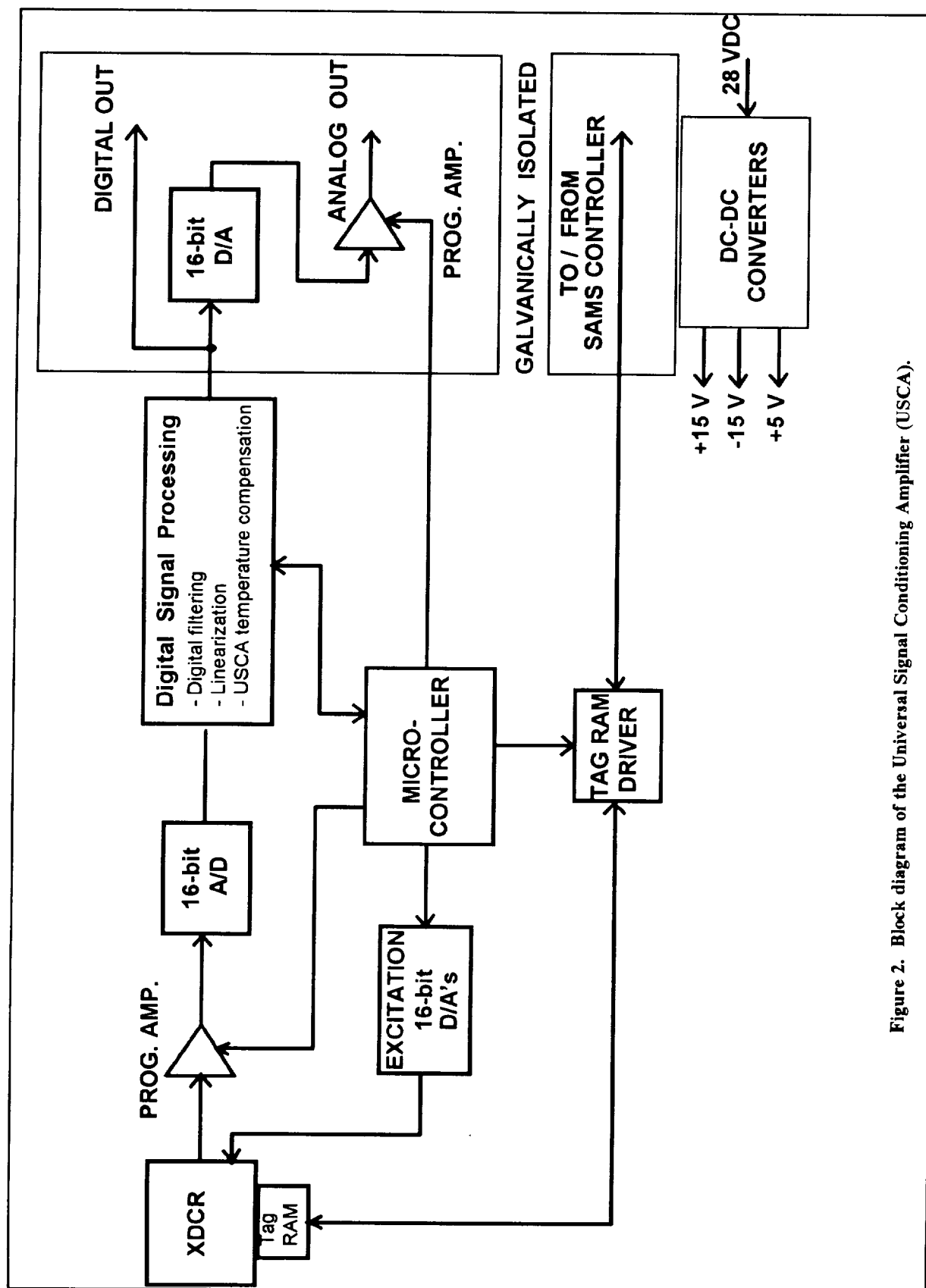


Figure 1. System configuration: Universal Signal Conditioning Amplifier (USCA), Tag Ram, and Self Aware Measurement System (SAMS)



The output module provides both analog and digital outputs. The analog capabilities were implemented to assure the compatibility of the USCAs with the measurement system currently in use at the Kennedy Space Center. A 16-bit digital-to-analog converter is used for the analog output. The output range can be selected to be 0 to 5 V, -5 to 5 V, -10 to +10 V, 0 to 10 V, or 4 to 20 mA. The digital output consists of a serial stream containing 16-bit data, thus the resolution is one part in 65,536. The output module is connected to the digital signal processing module by means of optoisolators. This assures a complete galvanical isolation between the output stage and the remaining sections of USCA.

USCA is powered by a 24-32 V source. The power supply module consisting of several DC-DC voltage converters provides the different voltages required for the operation of USCA. It also isolates the main circuitry from the output module, the data I/O module, and the power source.

A data input/output module is used to communicate with the SAMS controller and to read the information stored in the Tag Ram. Custom filters and new gain settings are downloaded through this module from the SAMS controller. This module is also galvanically isolated from the rest of the circuitry.

The operation of the USCA is controlled by a micro-controller. The information read from the Tag Ram is interpreted by the micro-controller, which uses it to define the settings for the operation of the USCA. This stage communicates with all the other modules. It sets the excitation voltages, the output range, selects the filters and gains, and continuously monitors the temperature inside the USCA. The temperature information is passed along to the digital signal processor, which uses it for the dynamic temperature compensation scheme. When the micro-controller detects that a transducer has been disconnected from the USCA, it will set both the excitation voltage or current and the output voltage to zero. This prevents the potential damage to another transducer when it is connected to the USCA and before the USCA reads its Tag Ram.

The micro-controller also stores the program for the digital signal processor chip. This program is transferred to the DSP upon powering the unit up. The DSP program can be upgraded remotely through the SAMS controller when required.

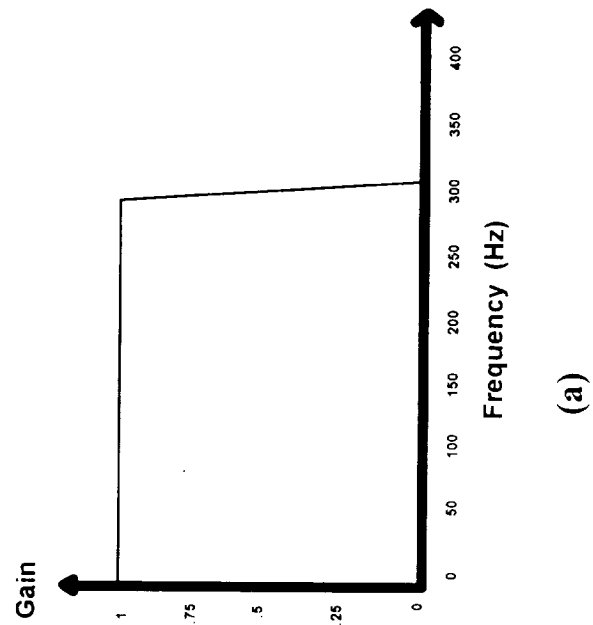
Digital Filters

A variety of digital filters were designed using Monarch DSP. The USCA micro-controller has a total of 32 kB of non-volatile memory, a section of which is used to store the 20-bit wide coefficients for the digital filters. The finite memory space limits the number of resident filters to eight. Figure 3 shows an example of two digital filters designed for USCA. Figure 3(a) depicts a 300 Hz lowpass filter and Figure 3(b) presents the same filter with a 60 Hz notch filter added to it. The passband ripple of these filters is 0.02 Db and the out-of-band attenuation is greater than 50 dB. Additional filters can be downloaded at any time using the SAMS controller. Filters can be customized to any given application. All the digital filters were designed using a Finite Impulse Response (FIR) implementation. FIR type filters provide linear phase delays, therefore better preserving the time domain waveforms.

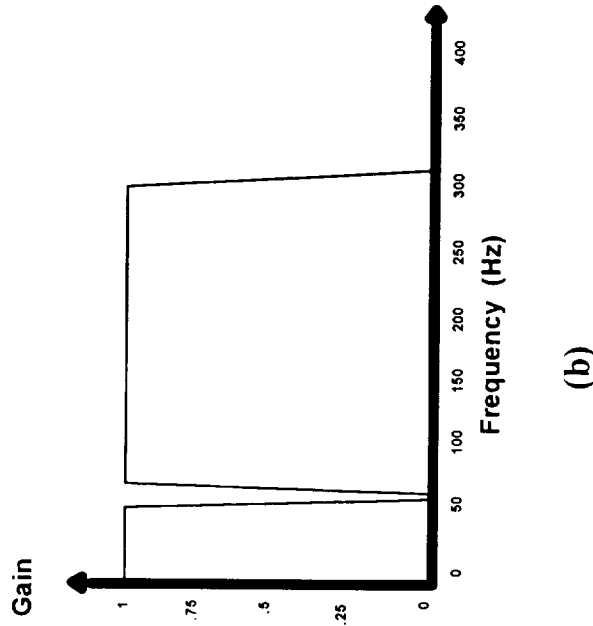
OPERATION

To utilize the USCA, a transducer must be provided with a Tag Ram containing the information pertinent to the transducer. To calibrate a transducer, its stimulus is varied and its output is recorded on a calibration data sheet. The linearization coefficients are then calculated and stored in the Tag Ram, along with information regarding the excitation required by the transducer. Information regarding the output range and the required gain/filter settings is also stored in the Tag Ram at that time.

When a new measurement is installed, the transducer equipped with a Tag Ram is connected to the USCA, without the need for any further calculations. The gain, filter, excitation and offset are automatically configured in a few seconds. The complete system configuration can then be reported by the SAMS controller.



(a)



(b)

Figure 3. Examples of digital filters generated for USCA.
 (a) 300 Hz lowpass filter.
 (b) 300 Hz lowpass filter with 60 Hz notch.

USCA DEVELOPMENT

A working prototype of USCA has been implemented at the Kennedy Space Center. The Tag Ram concept was developed and tested during 1991. A working USCA prototype and a 32-channel SAMS controller were implemented in 1992 and demonstrated in 1993.

The current work is concentrating on the miniaturization of USCA. New designs will be implemented using surface mount technology. Where available, surface mount components with a 50 or 25 mil pitch will be used. The miniature version of USCA will be placed in an explosion proof enclosure capable of withstanding the temperatures, pressure and vibration which are present at the launch pads when the Space Shuttle lifts off. To achieve better temperature stabilization, the USCA will be filled with a good thermal conductive fluid. The transient protection circuitry will be able to be replaced without opening the liquid filled portion of the environmental container. The calibration parameters for the USCA itself are stored internally on non-volatile RAM. A USCA can therefore be recalibrated by loading new calibration parameters, without the need of opening the liquid-filled case. Final characterization of the performance of USCA will be conducted following the completion of the miniature unit.

COMMERCIAL APPLICATIONS

The technology used in the development of the USCA can be utilized also in commercial and industrial applications. The USCA is an ideal solution for test cells, where sensors are frequently reconfigured for individual tests. The instantaneous matching of the USCA to a transducer can save several hours of measurement setup time. The ruggedized version of USCA can be used in situations where the environment is hostile, such as oil exploration and mining, where the USCAs would be subject to severe environmental conditions. A hermetically sealed USCA could be used when measurements need to be conducted under water. A simpler and less expensive USCA could be implemented for cases where large temperature variations are not expected and where the USCA would not be exposed to extreme shock or vibration. An interest has already been expressed on using the USCAs in wind tunnel measurements. We anticipate the use of the USCA in the automobile and aircraft industries, where multiple sensors are used during development and testing of components.

It is expected that within two years NASA will be able to procure USCAs from a commercial vendor. We anticipate the private sector will be interested in the commercialization of both ruggedized and laboratory-grade USCAs in the near future.

The Constant Current Loop: A New Paradigm for Resistance Signal Conditioning

541-32

2524

P. 13

Karl F. Anderson
Senior Measurement Systems Engineer
NASA Dryden Flight Research Facility
P.O. Box 273
Edwards, California 93523-0273

ABSTRACT

A practical single constant current loop circuit for the signal conditioning of variable-resistance transducers has been synthesized, analyzed, and demonstrated. The strain gage and the resistance temperature detector are examples of variable-resistance sensors. Lead wires connect variable-resistance sensors to remotely located signal-conditioning hardware. The presence of lead wires in the conventional Wheatstone bridge signal-conditioning circuit introduces undesired effects that reduce the quality of the data from the remote sensors. A practical approach is presented for suppressing essentially all lead wire resistance effects while indicating only the change in resistance value. Theoretical predictions supported by laboratory testing confirm the following features of the approach: (1) dc response; (2) the electrical output is unaffected by extremely large variations in the resistance of any or all lead wires; (3) the electrical output remains zero for no change in gage resistance; (4) the electrical output is inherently linear with respect to gage resistance change; (5) the sensitivity is double that of a Wheatstone bridge circuit; and (6) the same excitation wires can serve multiple independent gages. An adaptation of current loop circuit is presented that simultaneously provides an output signal voltage directly proportional to transducer resistance change and provides temperature information that is unaffected by transducer and lead wire resistance variations. These innovations are the subject of NASA patent applications.

NOMENCLATURE

dc	direct current
e_o	output voltage for Wheatstone bridge circuits, volts
E_x	excitation voltage for Wheatstone bridge circuits, volts
GF	gage factor
I	current level, amperes
I_A, I_B	directions of transducer current flow
M	measuring system for voltage difference
R	initial resistance of a transducer (strain gage, RTD, etc.), ohms
ΔR	change in gage resistance due to the quantity being sensed, ohms
R_{cal}	shunt calibration resistance, ohms
R_{ref}	Reference resistance, essentially equal to R , ohms
R_w	Lead wire resistance, ohms
RTD	resistance temperature detector
S	calibration switch

V	Voltage source for constant current regulation
V_A, V_B	Voltage outputs from I_A and I_B , respectively
V_g	Voltage across gage resistances
V_{out}	Voltage output from voltage difference measuring system
V_{ref}	Voltage across a reference resistance
$V+$	Positive thermoelectric (or other DC) voltage
$V-$	Negative thermoelectric (or other DC) voltage

INTRODUCTION

Variable-resistance detectors are convenient to use as transducers. Two common examples are the variable-resistance strain gage and the resistance temperature detector (RTD). All variable-resistance transducers change resistance in response to a change in the sensitive parameter. When the change in resistance is a small percentage of the total transducer resistance, the signal-conditioning task becomes more of a challenge. A strain gage is an example of a transducer that exhibits these characteristics. The significant output of a strain gage is the small resistance change, ΔR , that occurs in an initial resistance, R , as mechanical strain changes the dimensions of the gage. This small ΔR is typically no greater than 0.1 to 0.5 percent of R , which ranges typically from 120 to 1000 Ω .

The small magnitude of ΔR causes two general problems: the basic need to detect the small ΔR in the presence of the much larger initial gage resistance and the need to detect the desired ΔR in the presence of other parasitic resistance changes. Lead wire resistance changes, for example, can easily be larger than ΔR , especially where extreme temperature variations exist during a test. Serious errors due to lead wire resistance effects are common in the data from high-temperature strain measurements. High-temperature tests, such as those involving reentry simulations or operating turbine engines, often involve temperatures in excess of 1,000 °F. References 1 and 2 point out the continuing need to deal with lead wire resistance effects, especially where testing involves extreme temperature changes.

The Wheatstone bridge, in various circuit forms, is the traditional circuit used to signal condition variable-resistance transducers such as strain gages and RTD's. Although the Wheatstone bridge has been used effectively for many years in the signal-conditioning role, it has some disadvantages in many situations that can be overcome by a constant current loop approach. The constant current loop approach presented here is insensitive to any lead wire resistance changes, the basic cause of problems when the Wheatstone bridge is used as a resistance signal-conditioning circuit for single remote resistances. Diagrams and equations, along with test data from a practical circuit, are presented to illustrate the principle of operation. An approach for including several transducers (such as a strain-gage rosette) in the same constant current loop is presented. Each transducer in the loop has an independent output for indication and recording. Finally, an adaptation of this circuit is presented that simultaneously delivers both transducer resistance change and transducer temperature when thermocouple wire is used with a single-gage resistance in a four-wire circuit.

The key contributions of Allen R. Parker, Jr., who implemented the equations of the constant current loop signal-conditioning concept with practical circuitry and software, are gratefully acknowledged.

THE WHEATSTONE BRIDGE

A basic function of the Wheatstone bridge as a signal conditioner is to transform the small resistance change of the variable-resistance transducer into a proportional voltage signal. The Wheatstone bridge was originally presented in reference 3.

Background

The Wheatstone bridge does an excellent job of subtracting two large voltages to yield an output voltage due to ΔR , the change in resistance. The Wheatstone bridge output, however, is inherently nonlinear for resistance changes in a single arm, and current carrying lead wire resistances necessarily appear in the loop of the bridge circuit where they always reduce sensitivity and can cause zero shifts. Sir Charles Wheatstone states in reference 3, "Slight differences in the lengths and even in the tensions of the wires are sufficient to disturb the equilibrium [of my circuit]." Lead wire resistance effects are usually

manageable for tests that involve moderate changes in lead wire and connector resistances. Wheatstone bridge embodiments exist that reduce lead wire effects at the expense of added complexity in the signal-conditioning circuitry or additional lead wires or both. Reference 4 discusses using a Wheatstone bridge with three lead wires connecting to a remote strain-gage resistance in a way that minimizes the effects of lead wire resistance variations.

A variation of the Wheatstone bridge described in reference 5 deals with lead wire resistance problems by replacing two resistance arms in the bridge with constant current sources. A fourth lead wire is used, and a third constant current source is employed to force the current to be zero in what would otherwise be a current carrying lead wire. This tri-current method effectively tolerates lead wire resistance changes, delivers a linear output, and doubles sensitivity. Four lead wires, however, rather than three per measurement channel and three separate constant current sources that deliver very closely matched currents are required to obtain acceptable stability for strain measurements. The wiring, stability, and tracking requirements result in costs that have limited the application of the tri-current signal-conditioning technique.

Theory of Operation

The classic Wheatstone bridge circuit in figure 1 performs a precise analog subtraction of voltage drops across the various resistances in the loop of the bridge. This is done in a way that reliably isolates extremely small differences in large voltage drops. Commonly, microvolt (μV) level outputs are obtained from individual voltage drops of several volts. An ideal circuit is achieved for observing small variations in large resistances, especially when a minimum of parasitic lead wire resistance within the loop of the bridge is present.

The Wheatstone bridge circuit acts as an analog computation circuit. The excitation level serves as a multiplier to the circuit output, and variations in the four individual arms add to and subtract from the output according to their location in the circuit.

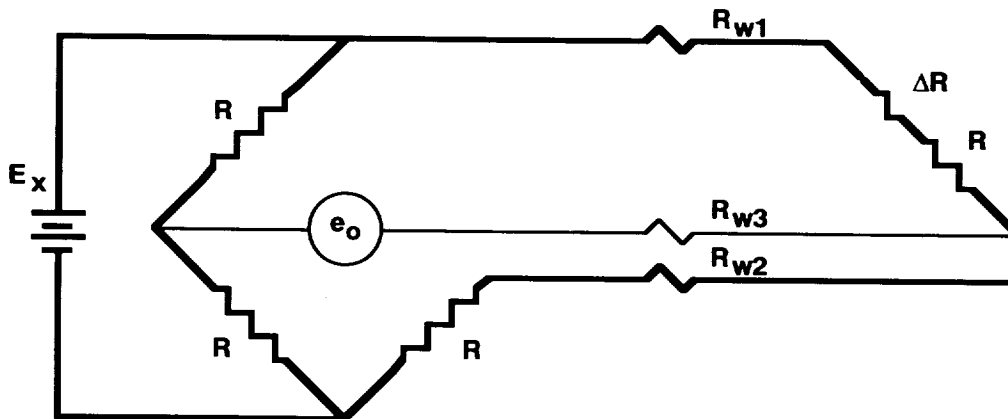


Figure 1. Single active arm Wheatstone bridge circuit.

The physical extension of the inner-bridge wiring of a Wheatstone bridge is often necessary in practice to measure the small change in a single remote resistance. In figure 1 these wires are modeled by their resistance, R_w . The lead wires to a remote bridge arm are arranged in the circuit of the figure so that their effects subtract from each other in the bridge output by taking advantage of the analog computation characteristics of the circuit.

These advantages account for the broad popularity and wide use of this circuit in most areas of electrical measurements, especially the measurement of small changes in large resistances. Strain-gage signal conditioning is but one of many transducer applications of the Wheatstone bridge.

Disadvantages

Some disadvantages, however, appear when employing the Wheatstone bridge circuit for signal conditioning a single remote resistance transducer. In this situation, lead wires become an uncontrolled contribution to the measurement output that appears as a variable systematic error.

The Wheatstone bridge circuit has a long history of effective use, extending back to the earliest precision measurement of electrical resistance. Its advantages normally outweigh its disadvantages by a wide margin in most resistance measurement applications. This situation is believed to account for the absence of developments to overcome its disadvantages, especially the difficulty with ill-behaved lead wire resistances.

The subtractive properties of the Wheatstone bridge require equivalent currents to flow in opposite directions through resistances connected to a common circuit node. Such a circuit arrangement always includes voltage drops due to currents in lead wires connecting the resistances to the common node in the voltages being subtracted.

Lead wires to remote bridge arms always appear within the loop of the bridge as additional, variable resistances that desensitize the system. Lead wire resistances must be identical within 0.2 milliohms (mΩ) for their effects to sufficiently subtract in the bridge output and thereby avoid an observable output drift that is indistinguishable from a change in the gage resistance. Lead wire resistances are uncontrolled in the usual test environment, and the resistance variation is much larger in test situations involving high temperatures.

A single active arm Wheatstone bridge circuit always delivers an output voltage that is a nonlinear function of the resistance change in that arm. And each independent single active arm measurement requires three wires to tolerate even mild lead wire resistance variations. The following three equations illustrate these effects based on the circuit of figure 1:

Equation 1 is the Wheatstone bridge output from a resistance change in a single arm with zero lead wire resistance.

$$e_o = \frac{E_x}{4} \left(\frac{\frac{(\Delta R)}{R}}{1 + \frac{(\Delta R)}{2R}} \right) \quad (1)$$

The appearance of ΔR in the denominator of equation 1 causes the output to be a nonlinear function of ΔR .

Equation 2 is the Wheatstone bridge output from a resistance change in a single arm with identical lead wire resistances, R_w , and additionally includes desensitization caused by identical lead wires.

$$e_o = \frac{E_x}{4} \left(\frac{\frac{\Delta R}{R}}{1 + \frac{R_w}{R} + \frac{\Delta R}{2R}} \right) \quad (2)$$

The appearance of R_w in the denominator of equation 2 causes the output to be a nonlinear function of both ΔR and R_w .

Equation 3 is the Wheatstone bridge output from a resistance change in a single arm with different lead wire resistances, R_{w1} and R_{w2} . This equation illustrates the major problem with lead wire resistance variation in high-temperature tests.

$$e_o = \frac{E_x}{4} \left(\frac{\frac{\Delta R + (R_{w1} - R_{w2})}{R}}{1 + \frac{(\Delta R + R_{w1} + R_{w2})}{2R}} \right) \quad (3)$$

The appearance of R_{w1} and R_{w2} in the numerator of equation 3 causes the output to be a direct function of their difference in addition to ΔR .

FOUR-TERMINAL VOLTAGE DIFFERENCE MEASUREMENT

The Wheatstone bridge uses a typical two-terminal voltmeter to indicate resistance changes. As discussed above, this approach has inherent limitations when lead wires are required in the bridge circuit

arrangement. A new measurement approach is to use a four-terminal voltage difference measurement system. This permits measuring a voltage difference without including the voltage drops across wiring to the common connection found in the Wheatstone bridge configuration. The four-terminal voltage difference measurement is at the heart of the improved measurement results obtained with the constant current loop conditioning concept described in the following section.

THE CONSTANT CURRENT LOOP

The dominant electrical property of a series circuit is that the current is the same through all parts of the series circuit loop. This property is very useful when long lead wires must be used in a measurement circuit. The ubiquitous 4–20 milliampere (mA) current loop in industrial process control is one practical example of using this property of series circuits. A strain-gage signal conditioning circuit using a constant current loop and a voltmeter circuit with the above voltage difference measurement characteristics is now in operational use at the NASA Dryden Thermostructures Research Facility.

Theory of Operation

Figure 2 diagrams the concept and illustrates the theory that explains its operation for a single-gage resistance. The unique part of the approach illustrated in figure 2 is the four-terminal voltage difference measuring system. R_{w1} through R_{w4} are lead wire resistances with R_{w1} and R_{w2} carrying the constant excitation current I . The gage is modeled by an initial resistance, R , in series with its resistance change, ΔR . Note that if the sensing system for the voltage across the gage, V_g , has a sufficiently high input impedance then no current will flow through R_{w3} and R_{w4} and, therefore, no voltage drop will occur across them. R_{ref} is a reference resistor used to develop a voltage, V_{ref} , which is subtracted from the voltage across the gage, V_g .

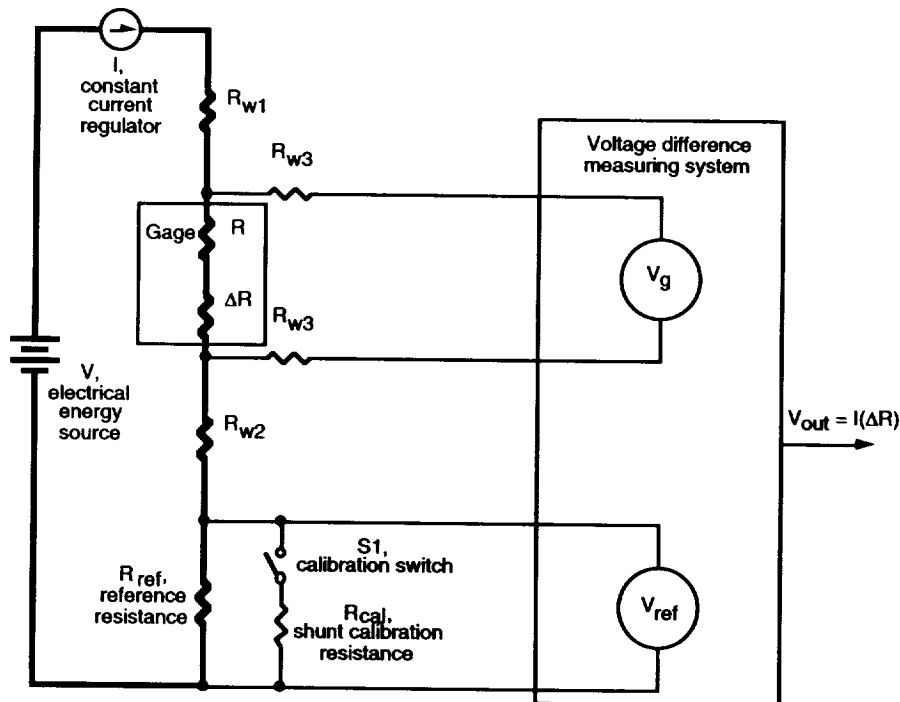


Figure 2. Current loop circuit for one-gage resistance.

The four-terminal voltage difference measuring system uses two terminals to sense V_g and two terminals to sense V_{ref} . Equations 4 through 6 model the circuit and illustrate the benefit of this four-terminal voltage measurement in a single constant current loop.

$$V_{out} = V_g - V_{ref} \quad (4)$$

$$V_{out} = I(R + \Delta R) - I(R_{ref}) \quad (5)$$

When $R_{ref} = R$,

$$V_{out} = I(\Delta R) \quad (6)$$

Note that R_w does not appear in the above equations.

When the voltage drop, V_{ref} , across a reference resistance in the current loop, R_{ref} , is subtracted from the voltage drop, V_g , caused by the same loop current I flowing through lead wires connected to a remote transducer resistance, R_g , the resulting output is equal to the difference in their resistances multiplied by the loop current. When their initial resistances are alike then the output is equal to the product of the current magnitude and the change in the transducing resistance, ΔR .

The high impedance of the voltmeter draws insignificant current through R_{w3} and R_{w4} , the voltage sensing lead wires, so virtually no voltage drop occurs along these wires to include in the circuit model. Therefore, no lead wire resistance appears in the equation for V_{out} , and the circuit is theoretically insensitive to any lead wire resistance changes. Measurements, discussed later, confirm that lead wire resistances play no significant role in the output from a practical circuit implementation.

A small difference between the initial gage resistance, R , and the reference resistor, R_{ref} will result in a correspondingly small output offset that can be subtracted out in data reduction as is also done with practical Wheatstone bridge circuits. This is a standard practice in strain-gage data reduction.

The maximum possible output voltage change per unit resistance change is achieved when using constant current excitation. By ignoring the second-order effects of the ΔR term in the denominator of the equation for the Wheatstone bridge output (equation 1), we have

$$e_o = \frac{E_x}{4} \left(\frac{\Delta R}{R} \right) \quad (7)$$

Since the E_x is $2V_g$ in a Wheatstone bridge circuit, the output in terms of the gage current and gage resistance change is

$$e_o = \frac{I(\Delta R)}{2} \quad (8)$$

Note that this is half the output available from equation 6, which describes the constant current loop output.

Advantages

The current loop has several major advantages as a resistance signal-conditioning circuit. The output bandwidth extends to dc. The loop performs a precise analog subtraction of only the desired voltage drops across the gage and reference resistances while ignoring the undesired voltage drops across lead wire resistances in the current loop. Analog subtraction is independent of the loop excitation current. Its output voltage is a linear function of the remote resistance change and is double what a single active arm Wheatstone bridge delivers for the same gage power dissipation. Large changes in various lead or connector resistances have essentially no effect on the output of a practical circuit. It is optimized for observing small variations in large resistances. And, as explained later, a strain-gage rosette system requires only six wires.

Disadvantages

A significant disadvantage of this circuit is that four (rather than three) lead wires are required to connect single remote gage resistances. Strain-gage rosette measurements require fewer wires, however, as discussed later.

Another disadvantage is that the circuit is not electrically balanced. The impedance from each end of the gage to test article ground is not identical. This can be a concern because the useful upper frequency may be limited by its common mode rejection characteristics. This noise can result from the conversion of

common mode energy to normal mode energy in the wiring from the gage to the signal-conditioning circuitry.

PRACTICAL CIRCUIT EXAMPLE

A practical means for accomplishing the function presented in figure 2 is illustrated in figure 3. The practical circuit in figure 3 uses a "flying capacitor multiplexer" circuit to subtract V_{ref} from V_g by transporting V_{ref} to another circuit location where it will directly subtract from V_g . This yields an output that is a function of only I and ΔR , because when the voltmeter draws no appreciable current, no lead wire resistances appear in the circuit equations.

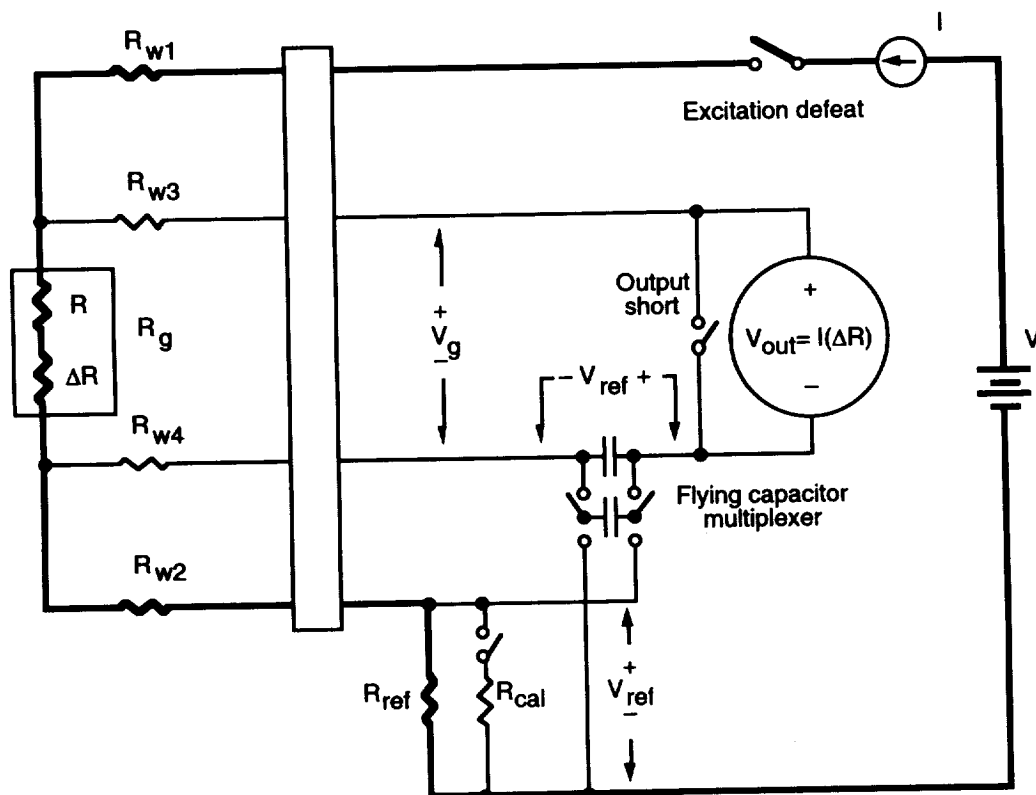


Figure 3. A practical current loop circuit for one remote gage resistance.

The flying capacitor multiplexer is only one of many possible means for implementing a four-terminal voltmeter that accomplishes the subtraction of two independent voltages regardless of what voltage exists between them. Commercially available switching components designed for flying capacitor multiplexers operate well in the up to 10-volt excitation levels needed for strain-gage work. Another advantage, discussed later, is the variety of analog computations possible in this manner, many that are beyond those available when using the Wheatstone bridge.

Circuit Features

The circuit in figure 3 includes a shunt calibration feature that operates in a manner equivalent to a shunt calibration with a Wheatstone bridge circuit. The resistance change that occurs as a result of paralleling R_{ref} momentarily with R_{cal} appears in V_{out} as though it were caused by a positive ΔR of the same magnitude.

Because the current is the same in all parts of the loop (indicated by heavy lines in all figures), there is no need to parallel R_g to achieve a useful calibration. And R_{ref} is precisely known while R_g is difficult to assess in a Wheatstone bridge circuit. So the constant current loop is capable of a simpler and more reliable direct resistance calibration than a Wheatstone bridge.

An excitation defeat function (sometimes called "power off zero") is included in figure 3 to identify any active noise that may be entering the system. With the power off zero activated, any extraneous noise from the installation environment may be identified and suppressed (if possible) to avoid contaminating the test data. The output short function is included in figure 3 to verify that no offset or common mode rejection problems exist in the voltmeter.

Circuit Component Requirements

The constant current source, I , should be stable to within less than 0.1 percent overall, have low output noise and, preferably, be both ohmically and electrostatically isolated from the power grid. As shown in equation 6, the magnitude of I directly affects the gain or sensitivity of the circuit. Constant current regulation is a common electronic function, and so no detailed circuit is presented here. An insulated gate field-effect transistor is appropriate to use as a pass element in the constant current regulator to achieve the highest practical output impedance.

Accuracy and stability requirements for the components in figure 3 are the same as for similar components in traditional Wheatstone bridge circuits. R_{ref} and R_{cal} are the same components found as bridge completion and calibration components in Wheatstone bridge circuits. Resistances stable to within 10 ppm/°F with 0.1 percent precision are normally used. Temperature stability is the more important characteristic since the magnitude of any resistance component is easily measured for use in data reduction.

The flying capacitor multiplexer switches in figure 3 are analog multiplexer switches using field-effect transistors. They are commercially available with timing and drive circuitry all integrated into the same package and designed for use as an instrumentation building block. Switch leakage current should be under 10 nanoamperes (nA), on resistance less than 1,000 Ω , and common mode rejection ratio greater than 100 dB.

The capacitors in the circuit are nominally 1.0 microfarad (μF) metalized polypropylene film devices with low dielectric leakage. They are a type commonly used for sample and hold purposes. The magnitude of capacitance does not need to be either precise or stable since their only function is to transfer and store electrical charge with minimal loss. Since the capacitors remain charged to constant levels in operation, no significant current surges occur in the circuit and the signal-to-noise ratio is high. It is usually not necessary to synchronize the flying capacitor multiplexer with other switching in the measurement system.

The calibration and output short switches are typically electromechanical switches or relays but field-effect transistor switches may be used if they have "on" resistances under 1 Ω and leakage currents under 10 nA. The calibration switch "on" resistance can be added to R_{cal} in calculating the response due to paralleling R_{cal} with R_{ref} .

The excitation defeat switch can be either an electromechanical relay or a transistor switch capable of carrying at least 30 mA. Power metal-oxide semiconductor field effect transistors (MOSFETs) having "on" resistances of under 1 Ω are suitable for this purpose. Excitation defeat can also be accomplished by programming the excitation current to zero.

The upper frequency limit of the circuit in figure 3 will be a function of the bandwidth of the constant current regulator, the bandwidth of the voltmeter (the voltage drop, V_{ref} , across R_{ref} is constant), and any electrical energy storage capability along the current loop.

Experimental Results

The circuit in figure 3 was used to gather data that demonstrates the sensitivity of its output to ΔR and its immunity to wide variations in the resistances of its lead wires. The tests involved extensive variations in both lead wire resistance (from 0 to 100 Ω) and gage resistance ΔR (from 0 to 5 Ω).

The following table summarizes the test results by presenting a limited set of data that cover the entire range of conditions tested. Data were reduced in terms of electrical resistance and microinches per inch ($\mu in/in$) of strain from a 120- Ω strain gage with a gage factor of 2. The ΔR is varied from 0 to 5 Ω in decades. The excitation current was 10 mA. The left column lists the input conditions in resistance change and equivalent microstrain (μs). The four right columns contain reduced data from voltage measurements using a bench top digital multimeter with 1.0 μV dc voltage and 1.0 m Ω resistance resolution. Offsets in the data set are all with respect to the initial indication with ΔR and all R_w values

set to zero. The bottom section of the table lists the measured R_w resistances in the various lead wires to identify the test condition for that column.

The data in the table show that the circuit output is a reliable function of ΔR . Note that, unlike a Wheatstone bridge circuit, the lead wire resistances R_{w1} through R_{w4} are not identical. While they appear to be closely matched, in fact the variation in resistance among the four lead wires exceeds the normal output from a 120- Ω gage resistance's ΔR of about 0.5 Ω . Such lead wire variations in a Wheatstone bridge circuit would render its output indications completely useless. But with the constant current loop circuit, *any* single lead wire can be varied by 100 Ω or more with insignificant variation in the output indication.

The tolerance of the circuit to lead resistance changes means that almost any wire size or connector can be effectively used, even slip rings. Lead wires can be chosen to be large enough to survive in the test environment rather than as large as practical to minimize the negative effects of lead wire resistance in a Wheatstone bridge circuit. This will result in smaller and lighter wire bundles and connectors between the sensors and signal-conditioning equipment.

Constant current loop signal-conditioning results reduced data for $R_{ref} = 120.0\Omega$.

ΔR	$\Delta R@R_w=0\Omega$	$\Delta R@R_w=10\Omega$	$\Delta R@R_w=30\Omega$	$\Delta R@R_w=100\Omega$
Ohms	Ohms	Ohms	Ohms	Ohms
0.00	0.000	-0.002	0.000	0.004
0.05	0.050	0.049	0.050	0.054
0.50	0.500	0.499	0.500	0.504
5.00	4.998	4.997	4.998	5.002
*Exact μs $\mu s = \Delta R / (GF \cdot R)$	$\mu s@R_w=0\Omega$	$\mu s@R_w=10\Omega$	$\mu s@R_w=30\Omega$	$\mu s@R_w=100\Omega$
$\mu in/in$	$\mu in/in$	$\mu in/in$	$\mu in/in$	$\mu in/in$
0	0	-6	0	16
208	208	204	209	225
2083	2084	2080	2086	2102
20833	20829	20827	20833	20849
Wire	Measured Lead Wire Resistance			
Identification	Ohms	Ohms	Ohms	Ohms
Rw1	0	11.237	29.113	100.256
Rw2	0	11.761	29.266	100.822
Rw3	0	11.351	29.400	100.375
Rw4	0	11.042	29.626	100.915

STRAIN-GAGE ROSETTE MEASUREMENTS

The same reference resistor voltage drop, V_{ref} , can be used as an input for more than one voltage difference function. This makes it practical to include more than one gage resistance in a single current loop with a corresponding reduction in the number of lead wires required.

Figure 4 illustrates three gage resistances, R_{g1} , R_{g2} , and R_{g3} , in a single loop. This configuration is applicable to the common technique of using a group of three strain gages installed near each other to estimate the magnitude and direction of principal strain. All the advantages of the constant current loop are obtained with only six lead wires, three fewer than required when using a Wheatstone bridge circuit for the same measurement requirement. This is made possible with a four-terminal voltage difference indicating system. One pair of terminals from each system is used to sense the voltage drop across its respective gage, and the second pair of terminals from each system is paralleled to sense the voltage drop across the single reference resistor in the constant current loop.

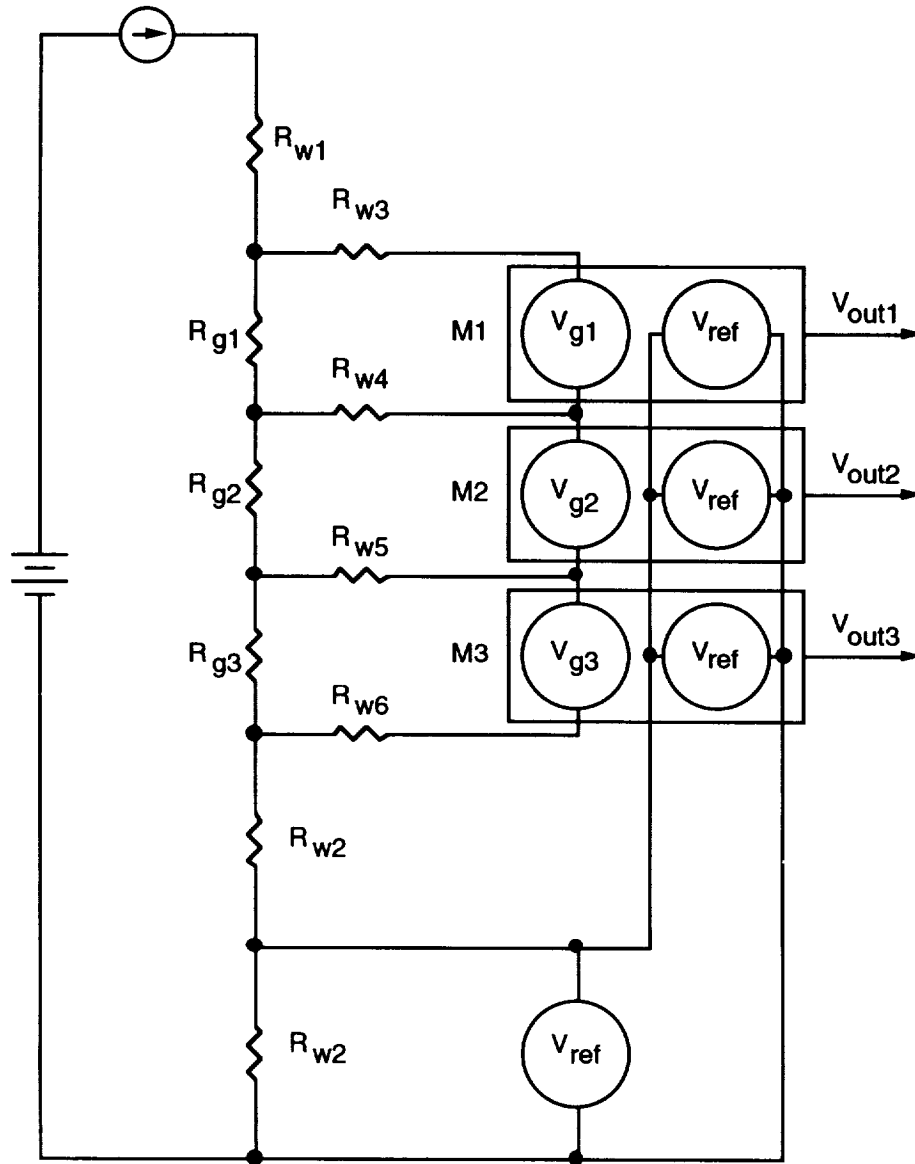


Figure 4. Strain-gage rosette measurement using current loop signal conditioning with six lead wires.

Figure 5 is a practical circuit employing the constant current loop in a strain-gage rosette measurement application. R_{w3} through R_{w6} are used to sense the three independent voltage drops across each gage resistance. The three output voltages are developed by subtracting the same reference voltage from each of the three gage voltage drops. The performance of this circuit for each of the three gage resistances is equivalent to the single-gage circuit. Observe that the quantity of lead wires is equal to three *plus* the number of gages in the loop. Rosette measurements using Wheatstone bridge circuitry require a quantity of lead wires equal to three *times* the number of gages.

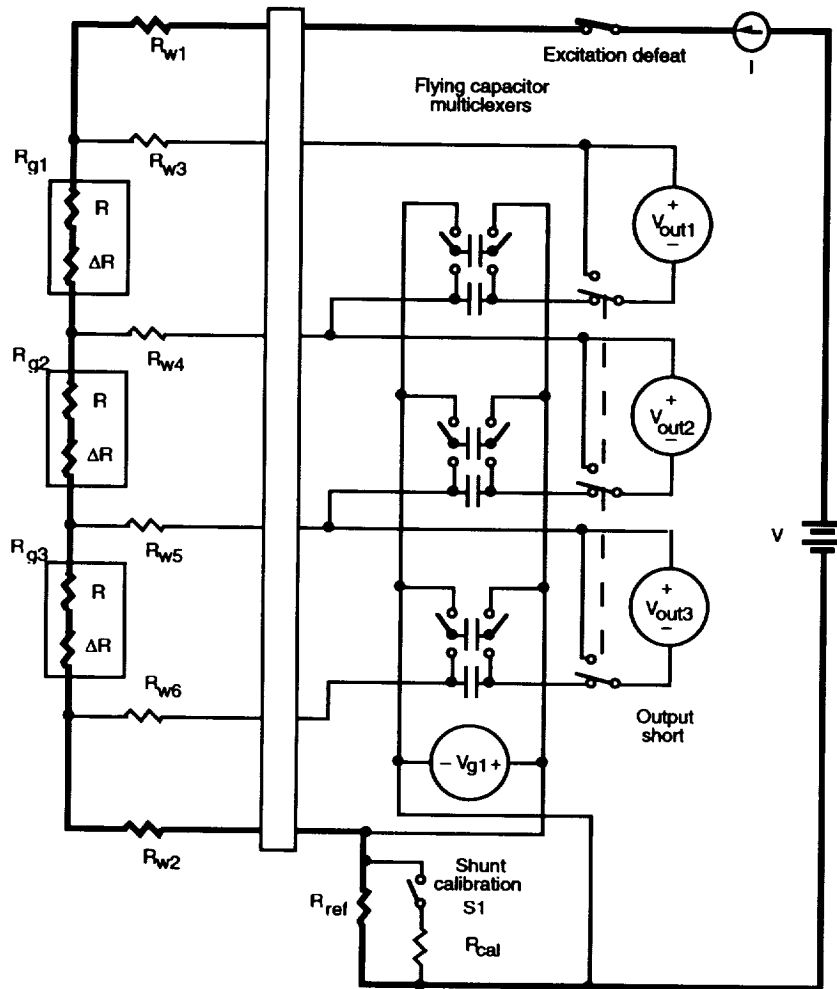


Figure 5. A practical strain-gage rosette measurement circuit using current loop signal conditioning with six lead wires.

The general approach discussed above can be adapted to other gage configurations. Apparent strain- and temperature-compensating resistances are often located near the gage resistance. The Hi Shear (ref. 6) and PdCr based high-temperature static strain gage (ref. 7) are examples of this need.

SIMULTANEOUS RESISTANCE AND TEMPERATURE MEASUREMENTS

Simultaneous resistance and temperature measurements can be made with an adaptation of the constant current loop circuit by using thermocouple wire for the lead wires that sense the voltage drop across a gage resistance. This is practical when a reversing rather than a unidirectional constant current excitation is used with appropriate signal processing.

An electronic double pole, double throw switch is used to reverse the current through a transducer that is provided by a dc constant current excitation source. By maintaining a constant current level while reversing excitation direction, as indicated in Figure 6, output voltages V_A and V_B are developed while the current is flowing through the gage resistance, R_g , as indicated by I_A and I_B respectively.

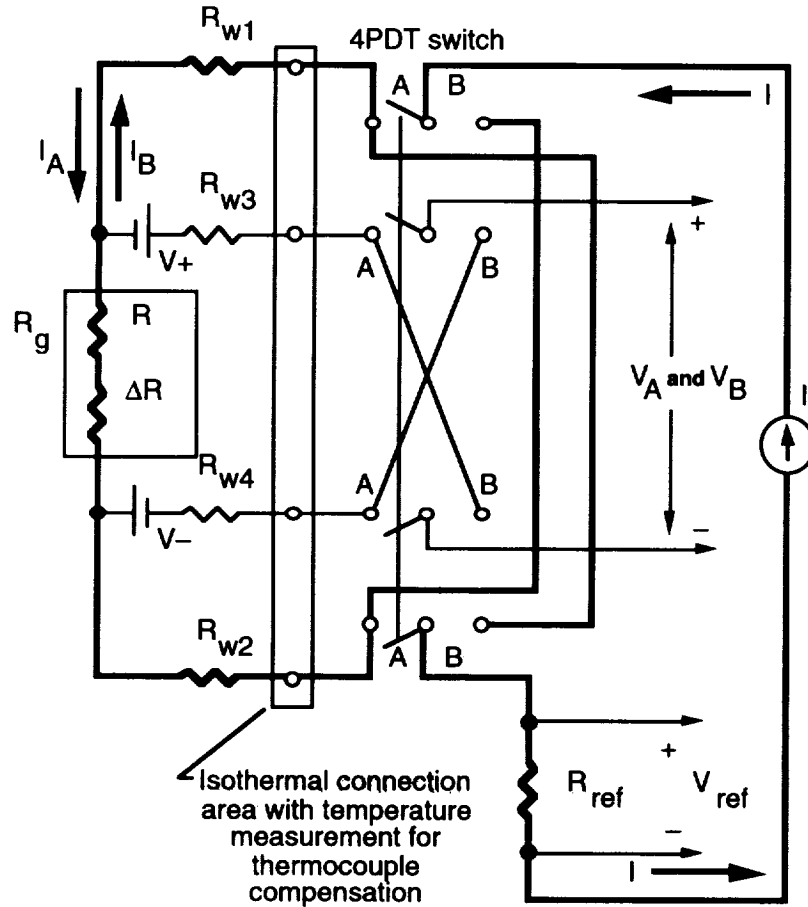


Figure 6. Reversing constant current loop signal conditioning, which separates resistance changes and voltage outputs into independent data channels.

When there is insignificant current drawn while measuring voltages, the circuit equations for figure 4 are

$$V_A = +[(V+) + (V-)] + I(R + \Delta R) \quad (9)$$

$$V_B = -[(V+) + (V-)] + I(R + \Delta R) \quad (10)$$

for $R_{ref} = 2R$,

$$V_A + V_B - V_{ref} = 2I(\Delta R) \quad (11)$$

which derives gage resistance change and

$$V_A - V_B = 2[(V+) + (V-)] \quad (12)$$

which derives the thermoelectric (or other DC) voltage.

Equations 9 through 12 show how V_A and V_B , the sensing outputs, can be processed to yield nonself-generating transducer resistance change uncontaminated by lead wire and thermoelectric (or other dc) effects. They can also be processed to yield self-generating thermoelectric (or other dc) effects uncontaminated by transducer and lead wire resistance changes. Flying capacitor multiplexers, synchronized to capture and sum the various signals, have been successfully used to accomplish the arithmetic processing.

In this manner, four lead wires, including a pair of thermocouple wires, provide an output proportional to ΔR uncontaminated by lead wire resistance and thermoelectric effects. A separate output is developed that is proportional to the temperature of the thermocouple lead wire attachments to the terminals of the gage resistance. This is accomplished while preserving all the advantages of the direct constant current version except frequency response.

Regardless of the capabilities of the sensors, system frequency response will be limited by sampling theory to no greater than half the excitation current reversal frequency. In practice, several samples will be required to demultiplex the output in response to an input change so the practical upper frequency response will be on the order of one-twentieth the excitation reversal frequency.

CONCLUSIONS

A practical signal-conditioning circuit for resistance transducers identified as the constant current loop has been synthesized, analyzed and demonstrated. Theoretical predictions and laboratory results agree, demonstrating that the output provides dc frequency response, is unaffected by extremely large variations in lead wire resistance, and is inherently linear. The sensitivity is double that which a Wheatstone bridge delivers for the same power dissipation in the gage resistance.

Fewer and smaller lead wires are needed in multiple transducer installations, such as strain-gage rosettes, than for the Wheatstone bridge. An approach has been described that separates self-generating (such as thermoelectric) and nonself-generating (resistance change) effects into independent data channels by using alternating constant current excitation for the loop along with appropriate signal processing.

REFERENCES

1. M.M. Lemcoe and Keith Krake, "Dryden Hi-Temperature Strain Measurement Systems Accomplishments and Future Work," Paper Number 263.B, NASP Mid Term Technology Review, Monterey, CA, April 21-24, 1992.
2. "NASP Strain Gage Workshop-1991," NASP Workshop Publication 1010, Proceedings of the First Nasp Strain Gage Workshop sponsored by NASA Lewis Research Center, Cleveland, OH, April 9-10, 1991. (Distribution restricted to U.S. government agencies and U.S. government agency contractors only. Other requestors should contact the NASP Joint Programs Office, Wright-Patterson AFB, OH.)
3. Wheatstone, Sir Charles, "An Account of Several New Instruments and Processes for Determining the Constants of a Voltaic Circuit," Philosophical Transactions of the Royal Society of London, vol. 133, 1843, pp. 303-329.
4. C.C. Perry and H.R. Lissner, *The Strain Gage Primer*, McGraw-Hill, Inc., New York, 1962
5. Barret B. Weeks and William E. Shoemaker, "Tri-Current Transducer Conditioner," ISA Preprint Number 9.1-3-65, 1965.
6. "Adverse Environment Weldable Strain Gages," Product Catalog, Eaton Corp., Troy Michigan, Oct. 1985.
7. J. Lei and W. Williams, "PdCr Based High Temperature Static Strain Gage," AIAA-90-5236, 1990

OMIT

VIDEO AND IMAGING TECHNOLOGY

LITERACITY: A MULTIMEDIA ADULT LITERACY PACKAGE COMBINING NASA TECHNOLOGY, RECURSIVE ID THEORY, AND AUTHENTIC INSTRUCTION THEORY

Jerry and Dee Anna Willis, Clare Walsh, Elizabeth Stephens,
Timothy Murphy, and Jerry Price
Center for Information Technology in Education
College of Education, University of Houston
Houston, TX 77204

342-85
2525
p. 10

William Stevens
Technology Utilization Office
Lockheed Engineering & Sciences Company
NASA, Lyndon B. Johnson Space Center, Houston, TX

Kevin Jackson
National Institute of Corrections
Department of Justice, Washington, D. C.

James A. Villareal and Bob Way
Advanced Software Architectures, Software Technology Branch
Information Systems Directorate
NASA, Lyndon B. Johnson Space Center, Houston, TX

ABSTRACT

An important part of NASA's mission involves the secondary application of its technologies in the public and private sectors. One current application under development is LiteraCity, a simulation-based instructional package for adults who do not have functional reading skills. Using fuzzy logic routines and other technologies developed by NASA's Information Systems Directorate and hypermedia sound, graphics, and animation technologies the project attempts to overcome the limited impact of adult literacy assessment and instruction by involving the adult in an interactive simulation of real-life literacy activities. The project uses a recursive instructional development model and authentic instruction theory.

This paper describes one component of a project to design, develop, and produce a series of computer-based, multimedia instructional packages. The packages are being developed for use in adult literacy programs, particularly in correctional education centers. They use the concepts of authentic instruction and authentic assessment to guide development. All the packages to be developed are instructional simulations. The first is a simulation of "finding a friend a job."

THE DESIGN PROCESS

To develop the instructional package we used a heavily modified version of a traditional instructional development (ID) procedure. We adapted the Four D instructional development model of Thiagarajan, Semmel, and Semmel [1], but it is very similar to models developed by Dick and Carey [2] and many others. This model has four stages: (1) Define, (2) Design, (3) Develop, and (4) Disseminate. Although this model does provide a framework within which to organize and manage the process of developing instructional materials, it has a decidedly behavioral and linear flavor. Because the instructional packages we are developing are based on cognitive models of learning and teaching, the relatively poor "fit" between the ID model we had adopted and the process we were actually using quickly became obvious. Traditional ID models seem most suitable when the goal is to develop traditional, text-focused tutorials based on relatively standard behavioral learning theory. A project that strays from tutorials or from behavioral learning theory may find many aspects of the traditional ID models become barriers rather than facilitators to smooth, effective development of materials. In this project, the book, *Learning With Interactive Multimedia: Developing and Using Multimedia Tools in Education*, by Sueann Ambron and Kristina Hooper [3] was a major inspiration for modifying and transforming the ID model we were using.

In the following sections, each of the stages in the Four D model is summarized, with extensions and revisions related to the development of multimedia materials based on cognitive instructional theories. Most of the extensions and revisions reflect two major conceptual shifts in the ID model we used: recursion and reflection. While the 4 D Model is linear, we used it recursively which reflects one aspect of a more cognitive model of learning as well as of instructional development. The steps in the Four D model are described below in sequence that implies they are stages or phases in the ID process which must be completed one after the other. As we used the model, the 4 Ds were not steps to be completed one after the other as if they were layers of bricks in a wall. They were, instead, tasks to be completed. Unlike each layer of bricks in a wall, the tasks in the process may be addressed many times as the project progresses. Design and Development is thus a recursive rather than a linear process.

The other major change in the Four D model involves reflection in action. Schon's work on reflective practice [4] emphasizes the need to think about and revise practice based on careful observation and analysis of what is happening in the practice environment. That approach, when applied to ID, places a heavy emphasis on obtaining feedback from students and instructors in the environment where the material will ultimately be used. In many ways, this represents an expansion of the formative evaluation procedures included in many behavioral ID models. The ID model we use in this project is thus a Recursive, Reflective Design and Development Model (R2D2).

THE DEFINITION TASK

In the traditional ID model this task includes the preliminary steps required to develop any type of instructional material. The purpose of Definition subtasks is the identification and definition of instructional requirements. The five subtasks in this stage include (1) front-end analysis, (2) learner analysis, (3) task analysis, (4) concept analysis, and (5) specifying instructional objectives.

Front End Analysis

The initial step in developing instructional material is an evaluation of the need. This process generally involves a specification of the need and an evaluation of existing materials to determine if needs are already being met. The front-end analysis for this project indicated some computer-supported instructional materials were available for adult literacy programs, but almost all of them were drill and practice or tutorial packages designed from a distinctly behavioral perspective.

Learner Analysis

This step traditionally involves developing a clear understanding of the target students. In this case, the primary users are adult literacy students. We treated learner analysis as an on-going process rather than a preliminary phase of development. We worked with a number of students at the Houston Read Commission, but the format was not a traditional learner analysis that involves assessment of learner skills. Instead, the students became part of the development team. We asked for input on interesting themes for the simulation, discussed topics that should be in the simulation, and received feedback on visual "look and feel" issues as well as instructional strategies the students preferred. Instructors at the Houston Read Commission also gave input on what their students needed, what they preferred, and ways the simulation could be made more interesting and relevant.

Task and Concept Analysis

After the need has been established and relevant student preferences, characteristics, and needs defined, the next step in a traditional ID model is to define the major skills to be acquired by the students and to analyze those skills. In traditional behaviorally oriented projects the focus is on breaking things down into subunits or components that can be taught separately. In reading, this approach leads to a discrete skills model of instruction that targets things like short vowel sounds and consonant digraphs. The result of this model in adult literacy programs is often relatively boring exercises that attempt to remediate weak underlying skills on the assumption that when the student develops those cognitive "muscles" reading will improve because reading is really a collection of skills. This model has not served adult literacy well and it may be one reason for the high dropout rate in many programs as well as the lack of enthusiasm shown by many participants who truly want to learn to read. In addition, many adults who do not have a functional level of reading ability have already completed and/or failed at many such exercises during their school years and have significant negative associations with them.

Using the R2D2 Model we approached task and concept analysis from a somewhat different perspective: authentic assessment and instruction. The primary goal was to deliver instruction and provide assessment directly related to authentic tasks an adult in this culture completes by reading. Those tasks are the focal point of the project and both instruction and assessment focuses on them. Thus, during instruction the task of the reader will be to accomplish a common task such as finding and applying for a job, locating and renting an apartment, using social services, or buying groceries. Assessment focuses on aspects of the reader's approach to the task that limit or prevent success. That is not to say that no discrete skills analysis or skills instruction is conducted or used. It simply means that it is framed within the context of how it relates to a particular task that is real or authentic for the student.

When task and concept analysis is considered within the framework of authentic assessment and authentic instruction, the analysis becomes less a job to be done once and for all at the beginning of the development process and more a continuous aspect of ID. For the project described in this paper the task was "Finding a Friend a Job" which involved selecting appropriate jobs, obtaining and completing the application, and submitting the application to the potential employer. The instructional product was built around the task. Actual instruction and help learning to read is available to the student when, and only when, the student requests it. The instruction and help given is determined by the student who selects it in order to complete the authentic task at hand.

Specifying Instructional Objectives

This is normally the final step in the first stage of a traditional ID model. It involves converting the results of the analysis of tasks and concepts into a set of objectives. They in turn provide the basis for designing the instructional package and developing evaluation and assessment strategies. As with task and concept analysis, specifying instructional objectives changes somewhat when viewed from the perspective of the R2D2 Model. In behavioral theory, specific instructional objectives such as "When the reader is faced with a paragraph of directions on how to drive from home to a potential employer's office, he or she draws a correct map based on those directions" are needed to guide the design process. In the R2D2 Model, which is based on authentic instruction and assessment, specific objectives evolve naturally out of the creation of a realistic instructional environment that represents a real world literacy task. Objectives keep designers focused and help ensure that the instruction leads to learning that transfers to the real world. It is not as important to write specific objectives first when the design involves creating an instructional version of some aspect of the real world. Concepts such as information landscapes, user support, and user interface become much more important than the development of a long list of specific instructional objectives.

THE DESIGN TASK

The focus of this task is the design of prototype instructional material. Two subtasks in most traditional ID models, Media Selection and Format Selection, are completed along with a third, Selection of a Development Environment. Two other traditional subtasks, Initial Design and Evaluation Strategy, are folded into other tasks rather than treated separately.

Media and Format Selection

The media we selected was interactive multimedia and the format selected was simulation. Because of the type of computer and media resources available, or affordable, in most adult literacy programs, we selected a single-screen simulation that can be run on a Macintosh or IBM computer with a CD-ROM drive and a sound system. The package puts the reader in the primary role in a simulation of an interesting and typical real world activity - finding a job. During the prototype development stage, the entire package, including sound effects, music, and speech files were stored in files on the computer's hard disk drive. The final version will be distributed on a CD-ROM that includes the computer program, including graphics, in track 1 with all the sound in CD-Audio format on tracks 2 through 99.

The "Finding a Friend a Job" scenario used in the first instructional package requires the reader to select a friend (Figure 1.), select a job from those advertised that fits the friend's needs and abilities (Figure 2.), and correctly complete a three part job application that serves as a model for the friend (The second part is shown in Figure 3.) . If the job selected matches the needs and talents of the friend, the simulation ends with a call from the friend who says "I got the job and I could not have done it without your help." If the job selected is

Figure 1.



Figure 2.



inappropriate the friend calls to tell the reader there was a problem and asks for more help. The simulation can then begin again.

Throughout the simulation, the reader can switch from simulation mode to help mode by clicking a button. When the reader does not know a word or phrase, he or she can ask for assistance. Readers click the problem word or words and then choose which of five different reading strategies they think would be most helpful. The five strategies, which are represented by 5 icons at the bottom of the help screens, are phonics, syntax, context, pronounce for the student, and read and explain. Figure 4. illustrates one step on phonic analysis of the word *duties*. For each strategy selected readers apply it with the guidance of spoken directions. Throughout the simulation, readers can ask for help by clicking a question mark button and replay verbal directions and explanations by clicking a speaker button on the top right of the screen. They can replay the voice version of printed directions, explanations, and information by clicking small speaker icons displayed beside the text. Spoken help, directions, and assistance is available in either Spanish or English.

Selecting a Development Environment

The core of the development environment was Authorware Professional on the Macintosh (APM) supplemented by C language subroutines as well as sound and graphics software including Photoshop, SoundEdit, SoundWave, and Canvas. Graphics, music clips, and sound effects were used from a number of clip collections including Killer Tracks music clips and Wraptides backgrounds. A major advantage of APM as the development platform is its ability to support a wide range of utility and support software and hardware. APM has provisions for integrating sound including music clips and voice from many sources, still images and video from CD-ROM and laser discs, and graphics from many paint and draw programs.

Authorware Professional for the Macintosh (APM) is widely recognized as a powerful development system. It is used by a number of NASA subcontractors, including Rockwell, and by Boeing. APM has a number of advantages over developing instructional packages in standard programming languages such as C. Perhaps the most important advantage for this project is the ease of learning factor. Experienced instructional designers can be "up and running" in Authorware in less than a week while estimates of the time it takes someone to become a proficient independent C programmer are as high as two to three years of full-time work and study. In addition, products developed on APM can be transferred to the Authorware Professional version for Windows (APW). APM's ease of learning, and ease of use, facilitates an important aspect of the development team's work pattern. In traditional ID models, each person takes a specialist's role. For example, a graphic artist may create illustrations, an instructional designer creates sequences or frames of instruction, and a content expert either creates or approves of the content included in the package.

In the R2D2 Model as used in this project, the team was composed of "specialists" who were primarily responsible for different aspects of the program such as graphics or sound. One member of the team, however, served as the designated programmer. The roles were, not, however, watertight compartments. Instead, almost all the team members worked some on all aspects of development. The content specialist, for example, wrote and revised some of the Authorware code and redesigned some of the icons. The sound specialist wrote some of the script and designed several of the help and instructional strategies. APM's ease of use made it possible for every team member to program. The interactive nature of APM made design recursion possible. Without a development environment like APM, reflection and recursion would be less useful, or discouragingly useless, because it is too difficult to make changes and adjustments to the program once it is "set in code." For example, instead of storyboarding an instructional sequence and then turning it over to the programmer to be written, the team could quickly answer many "what if" questions. What if we change the location of the icons, what if we add an explanatory screen between two existing screens, what if we change the layers a reader moves through to get help with an unknown word? In APM it is easy to set "Start" and "Stop" flags at points in the program, make changes in that section, and then run only that section to see how it will appear to students. The same feature makes it very easy to get student and teacher input and suggestions on different components as the program is evolving. For example, the team created four different "look and feel" designs and asked sixty adult literacy students to select the one they preferred. The same group listened to eight different musical styles and selected the one they liked the best. Such on-going and recursive feedback-revise cycles are much easier to complete using APM than regular programming languages.

Another prime advantage of APM is that its underlying language is C. This allows NASA's fuzzy logic, also written in C, to be easily interfaced with the created simulation and to analyze data produced when the

Figure 3.

Work History

Most Recent Employer: University of Houston

Address: Houston, Texas

Position:

Name of Supervisor

Dates of Employment: from to

Phone #:

Reason for leaving:

Salary or Pay Rate:

Describe in detail the work you did:

Military Experience:

Dates of Service:

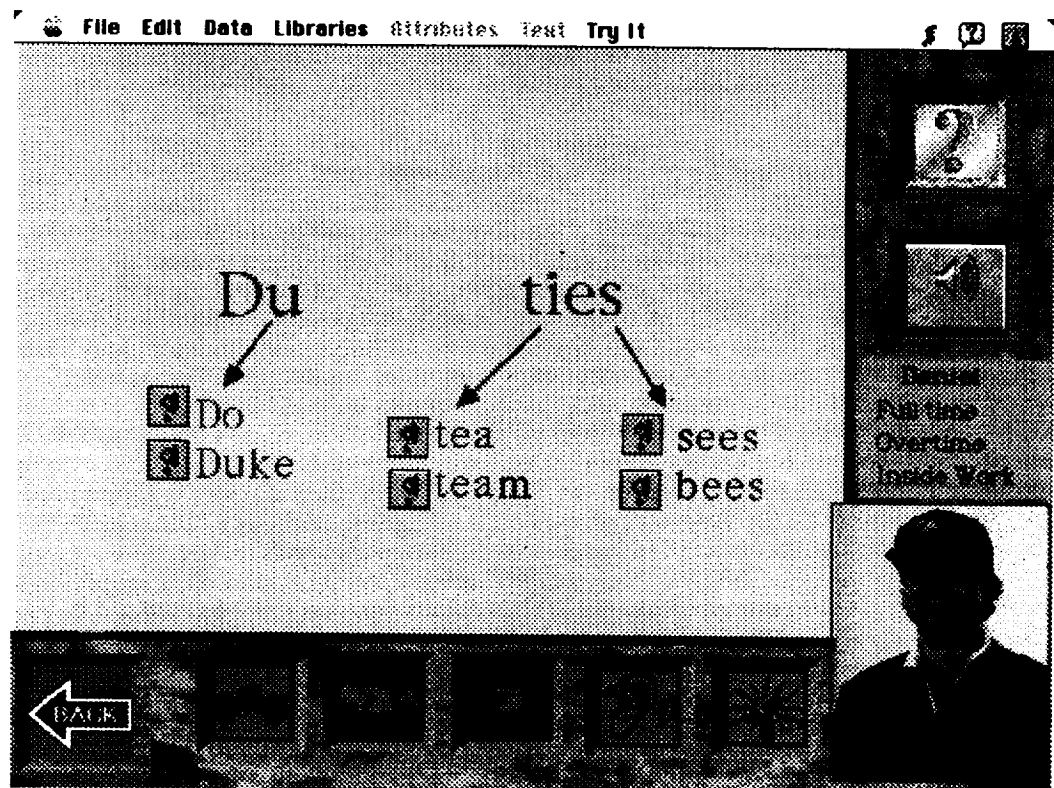
Type of Discharge:

Branch:

Date of Discharge:

BACK

Figure 4.



simulation is run. While the major emphasis of LiteraCity is placing control in the hands of the learner the data gathered, once analyzed and reduced to more manageable forms, can be of help to instructors guiding learners.

APM is not a perfect development environment, but it has much to offer when the ID model is recursive and reflective. An educational product is as much an artistic creation as it is a technical creation. We are all aware of technically correct educational materials, manuals, and documentation that numb the mind and discourage their use. This project aims to create a product that appeals to the user, one that has the snap and crackle that comes from well placed and well designed graphics, a product that has an appealing premise creatively executed. A traditional, top down, linear model of development is unlikely to produce such a product. What is needed is a development model that is the equivalent to the "management by walking around" movement in corporate leadership. That is, the designers and developers need to be able to "walk around" the program and try out possible alternative designs and arrangements with regular and ongoing feedback from potential users and consumers. The approach is easily implemented with Authorware. Changes, even major changes, are relatively easy to make while your program is running. And changes, once made, can be run immediately to view the effect from the perspective of the student who will be using the package. This advantage, which avoids the designer-to-programmer-to-software-to-designer cycle that has doomed so many instructional development projects, allows developers to "play with" many different alternatives in much the same way a musician can play with changes in a musical composition. This non-linear, real time approach supports improvisation, which leads to exciting designs. Eric Holsinger, in his article titled "Nonlinear systems enhance video editing" [5] makes the same point about electronic video editing. "Unlike standard video-editing systems, an editor can quickly cut and paste sections of the presentation together in any order. A user can save and compare different edits of the presentation easily, whereas starting a new version on a traditional off-line system requires you to start the edit from scratch each time."

Initial Design and Evaluation Strategy

The final two steps in the traditional Design stage are Initial Design and construction of an Evaluation Strategy. Because this project involves a product based on authentic instruction and authentic assessment theory there is less concern about a separate "test" of success. The simulation has a built-in assessment component and there is much less need for an external evaluation strategy such as pre- and posttests. The focus was thus on Initial Design. We approached the task of creating an instructional simulation by first developing an overall concept. Then we worked concurrently on three aspects of the general format and design:

1. Surface design - screen layout, typography, language, graphics, illustrations, sound.
2. Interface design - look and feel, user interaction, help, support, navigation, metaphors.
3. Scenario - sequence of simulation, options/choices, results.

We approached these three components by defining one path through the simulation that included every type of interaction, every type of help, and every type of outcome and feedback. We then completed all the work on that path so that the simulation ran from start to finish as planned - if one particular path was selected. At this point we have completed the surface design, interface design, and scenario for a single path. After students and teachers critique this path and revisions based on those critiques are completed, we will use the Authorware components in this path to build all the other paths. This reduces the effort required to develop the full simulation. One path serves as a template for the other paths through the simulation. In addition, many components in the Finding a Friend a Job simulation will be reused in the next simulation in the LiteraCity series.

THE DEVELOPMENT TASK

As the Design Tasks are completed, the Development Tasks become more and more important and relevant. An initial prototype is collaboratively developed and evaluated. Collaborative development involves the entire design team as well as consultants such as front line practitioners with recognized expertise and experience. As components of the prototype were developed they were subjected to two types of evaluation: expert appraisal and developmental tryouts. Both provide feedback that will be used to revise the prototype. In this project expert appraisal of the package was obtained from adult literacy practitioners and correctional education specialists. Developmental testing was carried out in Houston-area adult literacy programs.

We currently have a working prototype developed and before the program is ready for distribution the develop-get feedback-revise cycle will be completed many times for each component and several times for the complete prototype. Because Authorware is a flexible development environment it will be easy to make changes based on the feedback from both expert appraisal and developmental tryouts.

DISSEMINATION

Once a package has been revised based on feedback from expert appraisals and developmental tryouts, the final product is subjected to a summative evaluation. We have not reached this stage yet. The summative evaluation will involve using the package with a fresh group of students in an environment like the one where the material is to be used. Qualitative and quantitative data gathered from this evaluation will be included in the final teacher's manual included with the package.

The three Dissemination subtasks are Final Packaging, Diffusion, and Adoption. Final packaging involves creating and producing any necessary print materials (such as student guides and teacher's manuals), creating a master for the CD-ROM or laser disc and pressing a quantity of CD-ROM for initial use. If we are able to attract a commercial partner before final packaging, some of the work in this phase will be handed off to the partner. The commercial partner will also carry the primary responsibility for diffusion and adoption.

THE DESIGN AND DEVELOPMENT PROCESS

For this project a design team was organized. A wide range of specialists were required including programmers, instructional design specialists, content specialists and instructors, computer graphic artists, and project management leaders. The development of an effective team is not an easy task. There are, in fact, probably many more failures than successes in this field. The difficulty of the task is illustrated by Mark Heyer's article in the Microsoft Press book, *CD-ROM: The New Papyrus*. In his chapter Heyer, who has worked with Sony and Group W Westinghouse on CD-ROM and laser disc projects, describes what is hopefully a worst case scenario:

The conception, design, and production of visual/computer interactive programming is the newest challenge facing the creative community. . . .

In the videodisc industry, some organizations understand this problem [getting designers, visual producers, and programmers to work together] and have begun to create teams. In most cases, though, members of the three groups are separated in space and time. A few companies have even mandated that programmers will not talk to designers or video producers -- a clear failure mode.

Generally the process follows a path something like this: First a design document is generated and argued about for four to six months. So far it's a paper war. At best, 10 percent of the people who are judging and modifying the paper concept will actually understand the nature of interactive video.

Next the video is produced. Whether the content is motion or stills, this step normally consumes 80 percent of the total project time and budget. . . .

Then, after the visual material is mastered on a video or data tape, the control software is needed. In the case of videodiscs, most of which use external computer control, the disc is pressed and then the control software is written to conform with what has been put on the discs. Sadly, in many cases the programmer isn't even hired until after the disc is done. At that point the opportunity to make changes has passed. . . .

CD-ROM has an extra burden to consider, since the control software has to be pressed on the disc along with the visuals. Everything has to be perfect the moment the master is made. In reality, many producers will bear the expense of pressing discs with just the visuals as an aid to programming. This works about as well as any batch processing scheme, but it still doesn't allow for many visual change cycles, or for creative design input during programming. The computer programmer works with a fully interactive computer programming language and attempts to control a videodisc, which can be done in many ways, but the read-only pressed videodisc is in fact fixed and unchangeable.

The challenge of understanding these problems, and to some extent working around them, has occupied a whole generation of videodisc producers. We are just now in possession of design tools and methods which dramatically cut the development time and costs for videodiscs. The same techniques and equipment will work for CD-ROM [6].

Heyer paints a fairly accurate picture of many development efforts. Fortunately, he also offers some very useful suggestions. He suggests, for example, that design tools and systems must be developed and used that are relatively inexpensive and easy to use. Perhaps the most important point he made relative to this project relates to what he calls Interactive Editing. "For creative artists and designers, whether visual or computer, the ability to make unlimited small changes is an absolute benefit." Heyer suggests using a videodisc emulator so that "designers, visual artists, and programmers are working together on program creation in real time. Hundreds of changes can be made in a day, and each change is instantly demonstrable. Design and development time are cut to a fraction of what they used to be." Just such a recursive, real-time development process is essential in a project that depends on creativity for success. Heyer proposes that the design process begin with a crude storyboard that may be done on paper or on the computer using rough video and low fidelity graphics. Team members can work on the rough storyboard that may be done on paper or on the computer using rough video and low fidelity graphics. Team members can work on the rough storyboard, which is really a rough prototype, and experiment with a wide range of design options and alternatives. As the prototype matures, "real video and graphics are substituted for storyboard images, the code is finalized, and the project is finally sent off for mastering." Heyer sees three major advantages in this approach. Risk is reduced because fundamental problems can be discovered early. Because design and programming occurs at the same time, there is less likelihood that the team will commit to a design that cannot be executed. A second advantage is the encouragement of experimentation and revision which will enhance the quality of the final product. Finally the process of development becomes understandable and "each person knows exactly what the others are doing."

In this project we developed a storyboard (prototype) in Authorware Professional using "placeholder" graphics, photographs, sound and instructional sequences. Then, as the project progressed, placeholder materials were replaced with what we thought were final components. This was accomplished in two phases. First, a fully functioning path was developed through the entire simulation. Then the appropriate graphics, sound, and images were created and added to the program. That path was then evaluated by literacy teachers, literacy students, and experts in relevant areas such as graphics, instructional design, and assessment. Finally, the pattern and format that emerged from work on the path was extended to the other paths through the program. We are currently in that phase of the project: "filling in" the remaining paths through the simulation using the format, graphics, and structure already developed and critiqued by teachers and students.

A design and development approach that is recursive and allows for changes and revisions in "real time" calls for a collaborative team with a wide range of skills and expertise. It also calls for a team that can tolerate almost continuous change, adjustment, and fine tuning as well as the frustration that is probably inevitable when some work is abandoned or replaced by new material. The major drawback to traditional linear models of instructional development may be that they eliminate many opportunities for fine tuning and artistic enhancement. The major drawback to recursive models may be that everything is so fluid that team members have difficulty seeing progress and feeling good about their accomplishments to date. If everything can be changed at any stage of development, nothing is ever truly "finished" until the end of the project. Many team members need the feeling of closure that comes from finishing different stages of a long, complex project and teams that use a recursive model need to attend to those needs. [At the end of a long hard work session the "simplified-out" silliness that often occurs probably serves to alleviate the frustrations that have developed as well as build a sense of camaraderie.] Recursive approaches, like linear models, can be taken to extremes that are both frustrating and nonproductive. Whistler, for example, had great difficulty deciding when a painting was truly "finished." On several occasions he actually went to the homes of patrons who had purchased a painting and made small changes to it!

Non-linear, recursive development models must also deal with the fact that time, a factor in all projects, is linear. Almost all development projects have tight timelines, and recursive approaches call for nimble but energetic project management, particularly when the work of many different team members feeds into, and is required for, the work of others. Common problems in projects that use a linear development model -- assignments that are not completed on time, work that does not "fit" with work done by other team members, concepts and design formats that are changed by one group or individual but not communicated to others, and the "just a few more changes and it will be ready" phenomenon - can happen more often in recursive models and multiply the possibilities for problems, delays, and conflicts. Effective time management, long and short range project planning, successful team building, and strong overall project management are critical when a non-linear, recursive approach is used.

REFERENCES

1. Thiagarajan, S., Semmel, D., and Semmel, M. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Reston, VA: Council for Exceptional Children.
2. Dick, W., and Carey, L. (1985). *The systematic design of instruction (2nd ed.)*. Glenview, IL: Scott, Foresman.
3. Ambron, S., and Hooper, K. (1990). *Learning With Interactive Muldimedia: Developing and Using Multimedia Tools in Education*. Redmond, WA: Microsoft Press.
4. Schon, D. (1987). *Educating the reflective practitioner*. San Francisco: Jossey Bass.
5. Holsinger, E. (1992). Nonlinear systems enhance video editing, *MacWeek*, 10, 26.
6. Heyer, M. (1987). The creative challenge of CD ROM. In S. Lambert and S. Ropiequet (Eds.) *CD ROM The new papyrus*. 347-358.

**MAPPING ANALYSIS AND PLANNING SYSTEM
FOR THE
JOHN F. KENNEDY SPACE CENTER**

**C.R. Hall, M.J. Barkaszi, M.J. Provancha, N.A. Reddick, and C.R. Hinkle
The Bionetics Corporations
Kennedy Space Center, FL 32899**

**B.A. Engel
Agricultural Engineering Department
Purdue University
West Lafayette, IN 47906**

**B.R. Summerfield
Biomedical Operations and Research Office
Kennedy Space Center, FL 32899**

ABSTRACT

Environmental management, impact assessment, research and monitoring are multidisciplinary activities which are ideally suited to incorporate a multi-media approach to environmental problem solving. Geographic information systems (GIS), simulation models, neural networks and expert-system software are some of the advancing technologies being used for data management, query, analysis and display. At the 140,000 acre John F. Kennedy Space Center, the Advanced Software Technology group has been supporting development and implementation of a program that integrates these and other rapidly evolving hardware and software capabilities into a comprehensive Mapping, Analysis and Planning System (MAPS) based in a workstation/local area network environment.

An expert-system shell is being developed to link the various databases to guide users through the numerous stages of a facility siting and environmental assessment. The expert-system shell approach is appealing for its ease of data access by management-level decision makers while maintaining the involvement of the data specialists. This, as well as increased efficiency and accuracy in data analysis and report preparation, can benefit any organization involved in natural resource management.

INTRODUCTION

This paper presents an overview of a decision support system being developed at the John F. Kennedy Space Center (KSC) for use in environmental compliance, management and research. Solutions to environmental problems often involve complex, interdisciplinary subjects. Decision making must meet the requirements of such fields as engineering, hydrology, geology, ecology, geography, political science, public health, planning, demography and sociology [1]. Also, environmental management issues are by nature complex spatial and temporal phenomena. Their distribution across the landscape, in terms of quantity and quality, vary through time in response to both natural and anthropogenic factors. The pattern of landscape development (rural, agricultural, urban, industrial) historically evolved independently from environmental concerns and did not address the need to manage conflicts with natural resource protection, pollution abatement, or the quality of the human environment [2], [3].

Management of both landscape development and natural resources now requires a more diverse approach to conflict abatement. Such a pro-active approach is enhanced by using advanced computer-based technologies for information analysis, decision support, and visual display. Examples of these computer applications include but are not limited to multi-media (video, hypertext, sound, etc.), geographic information systems (maps, overlays, spatially distributed data, etc.), simulation modeling (surface and ground water, air quality,) remote sensing, expert-systems, and user friendly machine interfaces [4], [5].

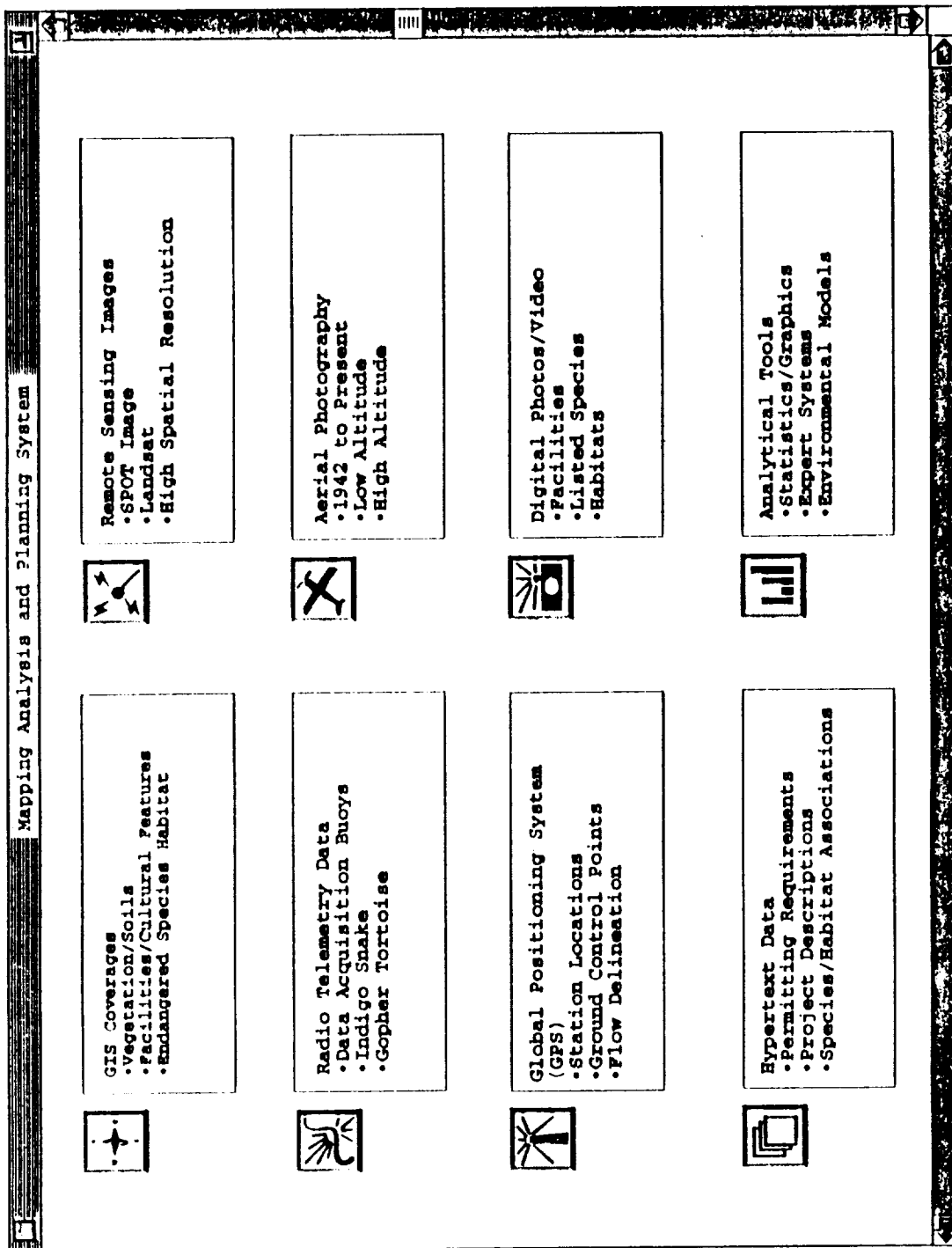


Figure 1. Schematic representation of the resources used in the development and implementation of the Mapping Analysis and Planning System.

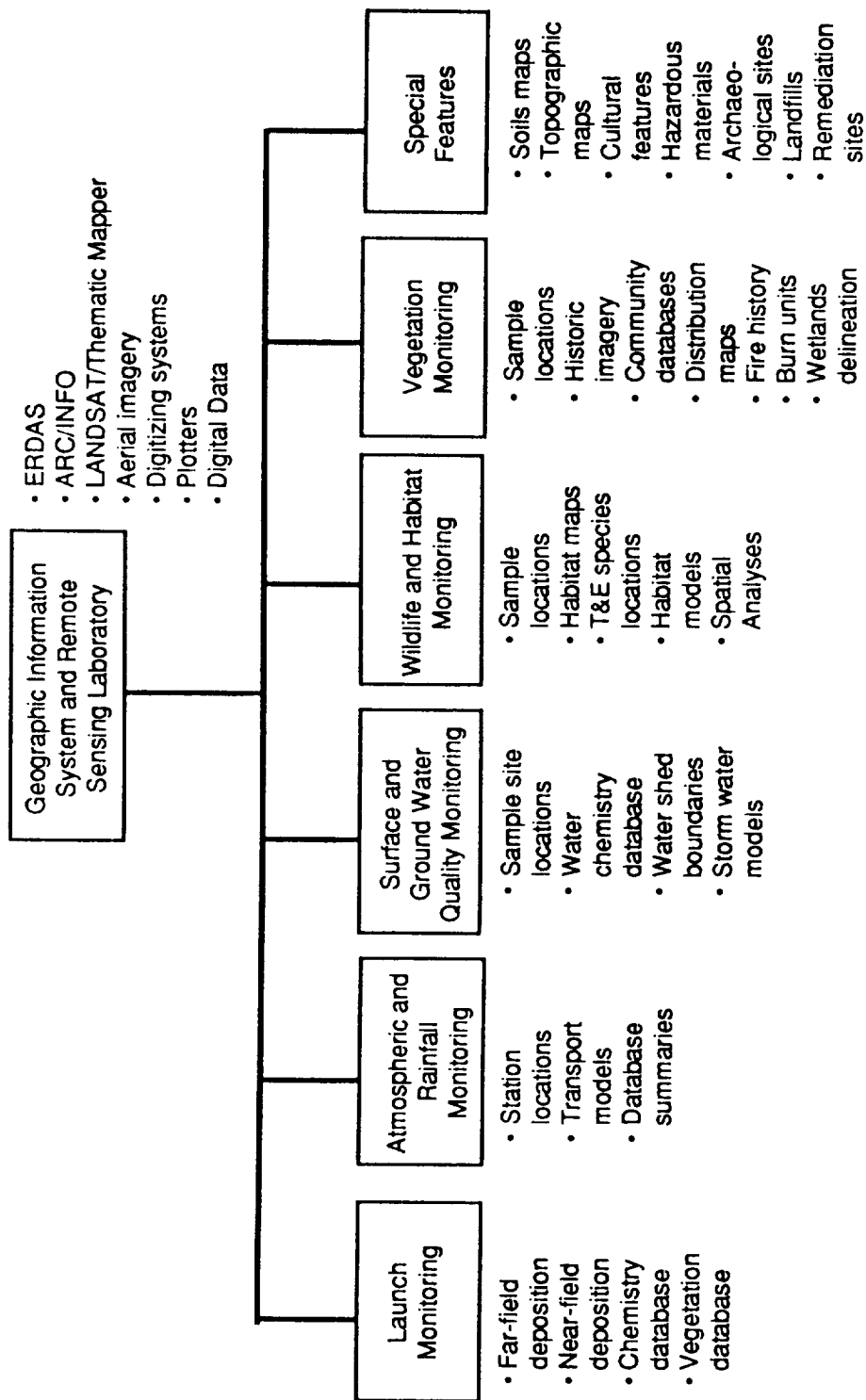


Figure 2. Long-term ecological data compiled in the GIS/Remote Sensing Laboratory

At the KSC, the Advanced Software Technology branch is working in conjunction with the Biomedical Operations and Research Office Pollution Control Officer to develop and implement a comprehensive, computer-based decision support system called the Mapping Analysis and Planning System (MAPS) which will be distributed across a network with the ultimate goal of making current information available to managers and planners at all levels. The MAPS will integrate the knowledge base from environmental research with policy, regulatory requirements, analytical tools and advanced display techniques (Figure 1).

Development of this concept and approach is currently a popular topic with many researchers and agencies exploring options in resource management. Results generally indicate that successful implementation of a comprehensive set of decision support tools requires that the users have available a large amount of high quality site specific data. Broad-spectrum data are needed regarding the site or landscape on which the system is directed as well as an understanding of the types of applications to which the system will be applied.

The primary database for the KSC MAPS consists of more than ten years of research involving field-collected and remotely sensed data covering a variety of subjects such as vegetation communities, wetlands, endangered species, habitats, soils, water quality, and cultural features [6] (Figure 2). Other site and project specific databases available within the MAPS include state and federal permits, permit monitoring data, identification of hazardous material storage areas, solid waste management units, landfills, space shuttle exhaust impacts, and facility siting constraints. Analytical capabilities include development of wildlife habitat association models, storm water and groundwater models and atmospheric diffusion models. Multi-media application include the ability to display digital imagery, captured video, audio, and hypertext to enhance the users understanding of environmental considerations, permit requirements, or regulatory constraints.

Both the environmental monitoring research database for KSC are in a continual state of collection and development. The KSC region is recognized as being biologically unique with more federally listed threatened and endangered species than any other protected area in the continental United States. The ongoing compilation of ecological data from such a diverse region make it some of the most current and reliable information available. In addition, the 140,000 acre facility is jointly managed by the U.S. Fish and Wildlife Service as the Merritt Island National Wildlife Refuge and the National Park Service as the Canaveral National Seashore.

In addition to the variety of sources providing an extensive monitoring and research database, environmental management responsibilities at KSC are also distributed across several NASA Offices and Contract organizations. Each has different responsibilities and information needs to meet requirements of state, and federal laws and regulations. Examples of laws and regulations that affect environmental management activities and generate various levels of information requirements include:

- Solid Waste Disposal Act
- Clear Air Act
- National Environmental Policy Act
- Resource Recovery Act
- Resource Conservation and Recovery Act
- Endangered Species Act
- Water Pollution Control Act
- National Energy Conservation Policy Act
- Pollution Prevention Act

Provisions of these laws are administered by a variety of Federal and State Agencies in Florida, including:

- Florida Department of Community Affairs
- Florida Department of Environmental Protection
- Management Districts
- Local Governments

United States Environmental Protection Agency
United States Army Corp of Engineers
United States Fish and Wildlife Service

Each of these organizations may require information or decision activities to design construction or operation of a facility for the purpose of minimizing economic or environmental risks and protecting human health. Stout and Streeter [7] state that the ultimate objective of laws and regulations is the minimization of potential risks associated with the activity of concern. Risk minimization is in turn best achieved by making informed decisions based on the best available data and information. The goal of the MAPS program is to coordinate the elements from all resources so that thoroughly informed decisions are more easily achieved.

APPROACH AND EXAMPLE

To combine regulatory requirements with other environmental information and data, the MAPS will link a variety of databases, digital imagery, GIS thematic layers, master planning files, video and legal information into a menu-driven decision support tool for KSC. A knowledge-based system shell is being developed to add expert site specific knowledge to the data fusion and query process. This will allow the users access to all available information, enhancing decision support and quality regarding environmental management. At present the system is being developed in a network environment utilizing UNIX-based work stations and DOS based personal computers. Software being incorporated into the system is a combination of commercially available packages. Basics of the system include:

- GIS software
- Image Processing software
- Video Capture and Editing
- Statistical Data Analysis
- 3-D Graphics
- Numerical Modeling
- Hypertext software
- Expert-systems/Neural Net Development packages

The system is currently being used for facility sitings where concerns must be addressed following requirements of the KSC environmental checklist [8]. The following paragraphs give a description of how the MAPS is used in the decision making processes for the siting of a new fuel storage facility.

The environmental checklist is the precursor to decision making for the siting of any facilities on KSC (Figure 3). Responses to questions on the checklist will either flag ancillary requirements, offer more detailed information on a topic, or prompt the user to input additional data. For example, a fuel storage facility will require a "yes" to the hazardous material storage question. The user is then prompted to supply the type of hazardous material as well as the amount. This system will then alert the user of any permits or official actions such as requirements for public meetings which need to be taken for that substance.

Depending on the question, there are currently different levels of maps, hypertext information, video and graphics to further the query by the user. At this time only some of these topics offer additional exploration; however, as the database size increases so will the extent of each checklist topic. Any topic in the checklist for which the user cannot provide a definitive answer must be checked "yes"; thus prompting the user to gather more information through the query process that will enable them to correctly respond. For the fuel storage facility example, the user may be unsure if the facility will be build in threatened or endangered species habitat; therefore, the user must check yes. The system then prompts the user to identify which habitats are present or, based on the location and surrounding area, defines the habitats for the user. Using GIS vegetation maps, the user finds that the proposed site is located in "scrub & slash pine" and "broadleaved woodlands". For each of these habitat types, the user may query hypertext or graphics for the following information:

5. DOES THE CONSTRUCTION, INSTALLATION, REMOVAL, ACTIVATION, OR OPERATION OF THE PROPOSED PROJECT INVOLVE: (FILL OUT ONE PAGE FOR EACH ALTERNATIVE CONSIDERED: SEE INSTRUCTIONS. WHEN IN DOUBT, MARK YES)

ALTERNATIVE # AND DESCRIPTION:

YES	NO	
<input type="checkbox"/>	<input type="checkbox"/>	a. Discharge of any substance to the environment Mark all appropriate media: list substance(s) in Block 6 Air <input type="checkbox"/> Surface Water <input type="checkbox"/> Groundwater <input type="checkbox"/> Soil <input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	b. Land alteration, excavation or dewatering
<input type="checkbox"/>	<input type="checkbox"/>	c. Construction in wetlands
<input type="checkbox"/>	<input type="checkbox"/>	d. Construction in floodplain If Yes: 100 Year <input type="checkbox"/> 500 Year <input type="checkbox"/> (Mark both if appropriate)
<input type="checkbox"/>	<input type="checkbox"/>	e. Generation of ionizing or non-ionizing radiation or use of any radiation source
<input type="checkbox"/>	<input type="checkbox"/>	f. Asbestos-containing materials or facilities
<input type="checkbox"/>	<input type="checkbox"/>	g. PCB-contaminated materials or equipment
<input type="checkbox"/>	<input type="checkbox"/>	h. Generation of waste other than normal construction debris If Yes, list waste (s) in Block 6
<input type="checkbox"/>	<input type="checkbox"/>	i. Use or storage of Hazardous or Toxic Materials If Yes, list materials and quantities for each in Block 6
<input type="checkbox"/>	<input type="checkbox"/>	j. Aboveground or underground storage tanks If Yes, list material (s) stored in Block 6
<input type="checkbox"/>	<input type="checkbox"/>	k. Generation of high noise levels outdoors (above 85 dBA)
<input type="checkbox"/>	<input type="checkbox"/>	l. An area of archaeological significance If Yes, indicate potential: High <input type="checkbox"/> Medium <input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	m. Endangered species habitat
<input type="checkbox"/>	<input type="checkbox"/>	n. Solid Waste Management Unit (SWMU) site
<input type="checkbox"/>	<input type="checkbox"/>	o. Other issues which could produce environmental impacts If Yes, describe in Block 6

Figure 3. Environmental checklist used by NASA/KSC resource managers and planners.

VEGETATION AND STRUCTURES

Launch complexes 39A, 39B and VAB area



Figure 4. Example of the KSC vegetation map, with existing structures overlay, that can be accessed from the MAPS

Amphibians and Reptiles	FGFWS	USFWS	CITES	FCREPA	Global Rank	FNAI State Rank
American Alligator	SSC	T (S/A)	II	SSC	G5	S4
Loggerhead Sea Turtle	T	T	I	T	G3	S2
Atlantic Green Turtle	E	E		E	G3	S2
Eastern Diamondback Rattlesnake						
Leatherback Turtle	E	E	I	R	G3	S2
Eastern Indigo Snake	T	T		SSC	G4T3	S3
Atlantic Hawksbill Turtle	E	E	I	E	G3	S1
Gopher Tortoise	SSC	UR2		T	G3	S3
McLe Kingsnake				R	G5	S4S3
Eastern Kingsnake*						
Atlantic Ridley Turtle	E	E	I	E		
Eastern Coachwhip*						
Florida East Coast Terrapin*						
Atlantic Salt Marsh Water Snake	T	T		E	G5T1Q	S1
Florida Pine Snake	SSC	UR2			G5T1Q	S?
Gopher Frog	SSC	UR2		R	G4	S3
Florida Scrub Lizard						
Dusky Pigmy Rattlesnake*						
Florida Crowned Snake*						
Coastal Dunes Crowned Snake*						
Birds						
Cooper's Hawk				SSC		
Mottled Duck*						
Bachman's Sparrow		UR2		R	G3	S?
Roseate Spoonbill	SSC				G5	S2S3
Dusky Seaside Sparrow	E				G4TX	SX
Florida Scrub Jay	T	T			G5T3	S3
Limpkin	SSC			SSC	G5	S3
Burrowing Owl	SSC		II	SSC	G5T3	S3
American Bittern*						

Figure 5. Example of chart used in maps showing the listing status of species found on KSC

Breeding Season	Seasonal Abundance on KSC				Estimated Population Size on KSC	Autecology Characteristics (Noss and Labisky in press)
	Spring	Summer	Fall	Winter		
Florida Scrub Jay	C	C	C	C	2,100	1 R ₁ D L U F H
Atlantic Green Turtle	C	O	C	C	150	
West Indian Manatee	C	C	C	R	< 300	
Southeastern Beach Mouse	C	C	C	C	> 5000	
Southern Bald Eagle	U	U	U	U	20	2/3 R ₂ W L C
Wood Stork	C	C	C	C	0-600	
Eastern Indigo Snake	C	C	C	C	> 5000	
Roseate Spoonbill	U	C	U	R	400	
Reddish Egret	O	O	O	O	20	
Florida Long-tailed Weasel	R	R	R	R	< 500	2 R ₁ R ₂ W L
Atlantic Salt Marsh Snake	R	R	R	R	< 500	
Florida Pine Snake	R	R	R	R	< 500	2/3 R ₂ N F H E
Florida East Coast Terrapin	R	R	R	R	400-500	
Atlantic Loggerhead Turtle	L	C			1,200	

Figure 6. Example of chart used in the maps showing the breeding season , population status and ecological characteristics of species found on KSC.

PRIMARY & SECONDARY SCRUB JAY HABITAT ON KSC

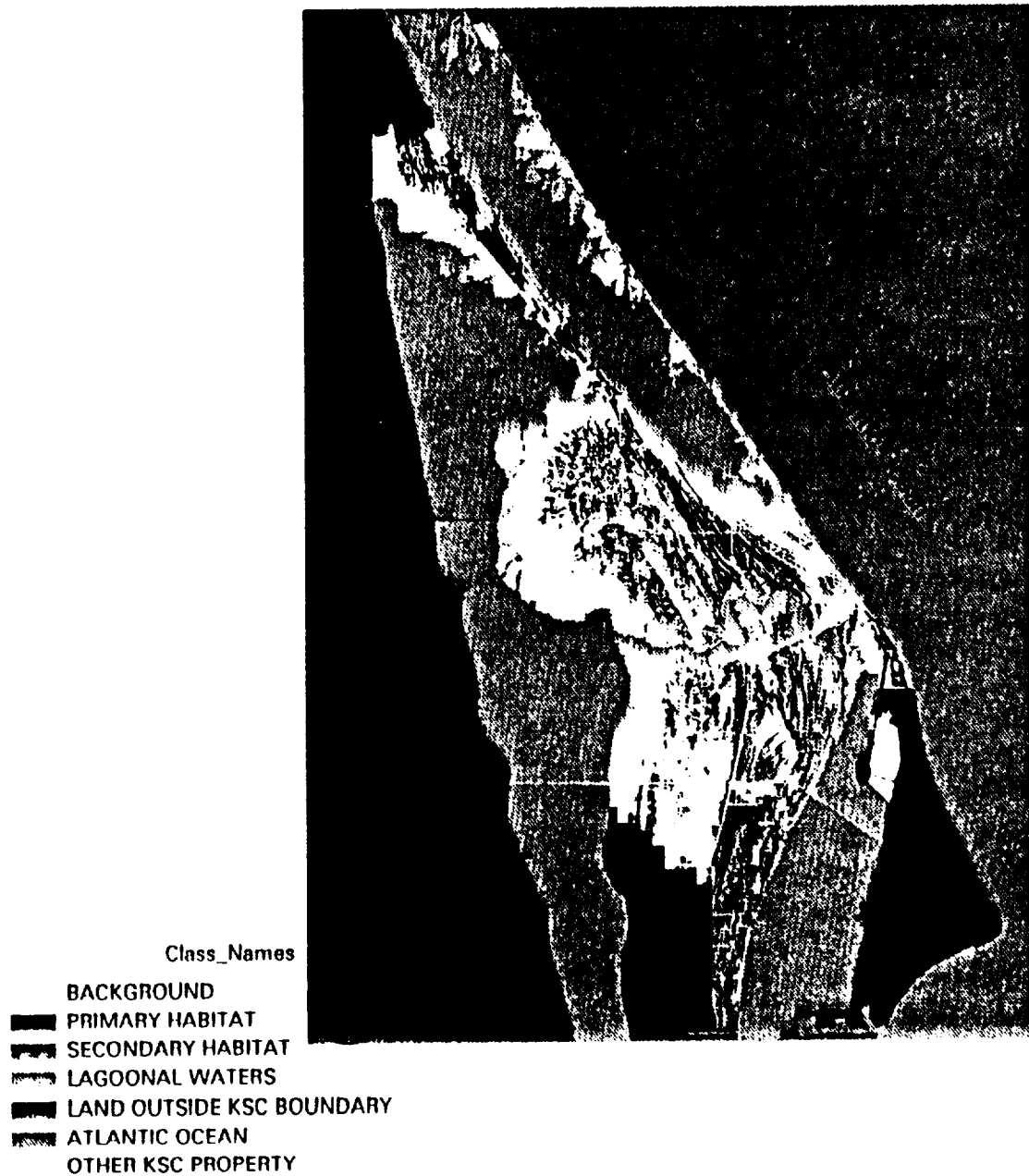


Figure 7. Example of species-specific information that can be accessed from the MAPS. This map shows the locations of primary and secondary habitat for the Florida Scrub Jay on KSC.

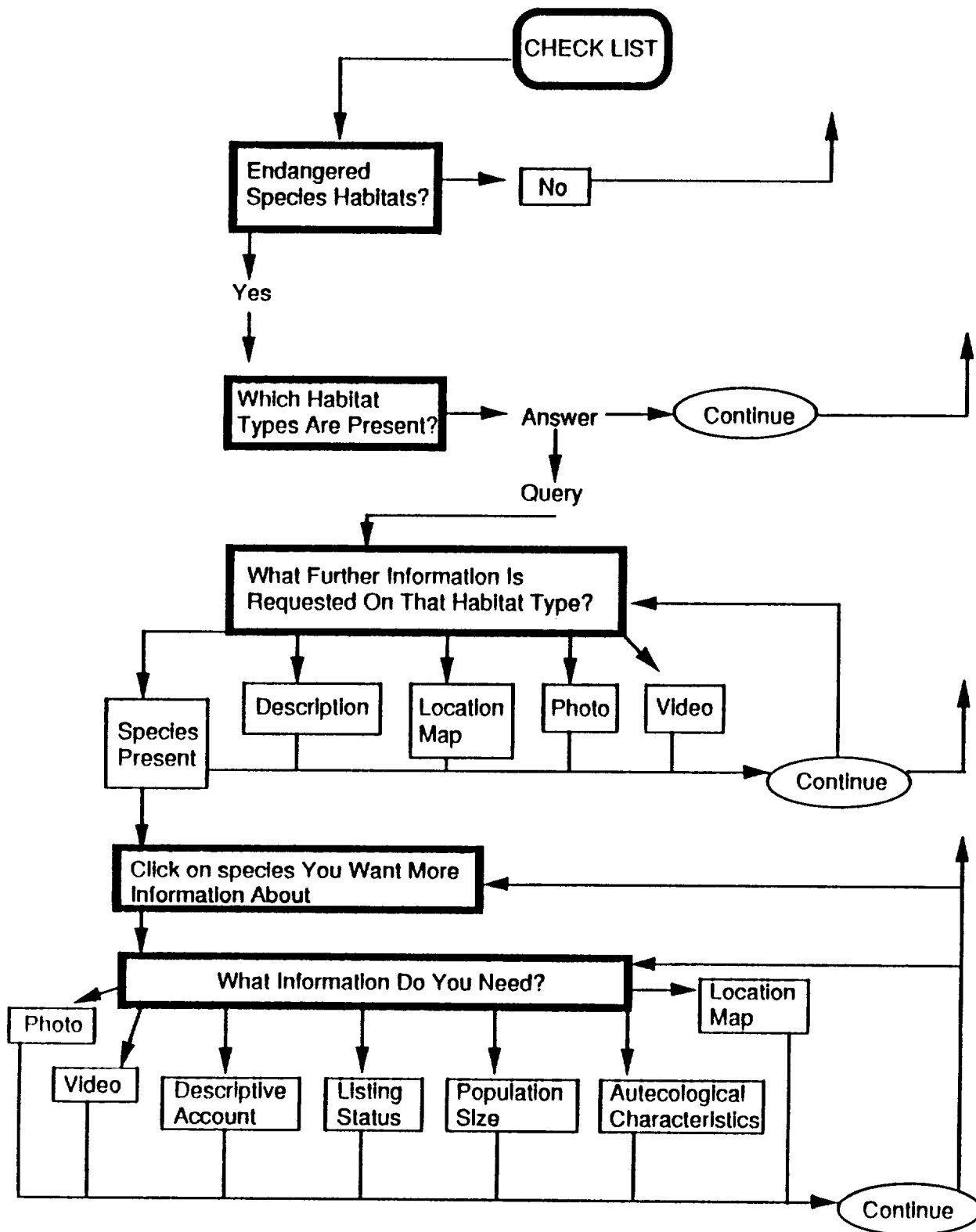


Figure 8. Flowchart of query process for endangered species habitat

- 1) A written description of the habitat. This description will include physical parameters, historical view points, plant and animal species typical for that habitat and NASA's influence on the habitat type within KSC.
- 2) A map showing the locations of that habitat on KSC (Figure 4). Further links with ARC/INFO and ArcView will provide access to finer scale maps of the site or LANDSAT and SPOT images. It will also enable the user to manipulate the map in order to query more information such as getting geographic coordinates or producing a measured radius from a site to look for possible contamination impacts.
- 3) Photographs and video segments of the habitat type.
- 4) A list of species found in that habitat. Each species can then be further explored to provide the user with detailed information on the ecology of each species.
- 5) The protected status of each species for five main listing agencies: U.S. Fish and Wildlife Service (USFWS), Convention on International Trade of Endangered Species (CITES), Florida Game and Fresh Water Fish Commission (FGFWFC), Florida Committee on Rare and Endangered Plants and Animals (FCREPA), and the Florida Natural Areas Inventory (FNAI) (Figure 5).
- 6) The species population size, breeding season, and seasonal abundance on KSC (Figure 6).
- 7) A map showing the locations on KSC where appropriate habitat for supporting that species is located (Figure 7).
- 8) A detailed, written account of each species which includes a description of the species and its typical habitat, listing status, reasons for endangerment, discussion of the species status on KSC, research results, behavioral notes, and general ecological attributes.
- 9) Photographs of the species.
- 10) A video segment of the species.

The flowchart in Figure 8 show a summary example of how a topic can be explored. One of the main advantages of this program is its ability to adapt to different user's requirements. A user can go through the query only as far as they need. For instance, not all users will need to know what certain species or habitats look like while others may need a photograph or video to confirm sightings. More users may need the listing status or population size for a species and the location of certain habitats on KSC. The user drives which information is presented. The available database can be readily updated as more information becomes available or as changes occur in legal status such as species' listings or permitting requirements.

SUMMARY

This brief paper outlines several of the development strategies and the status of the MAPS decision support process for environmental management issues at KSC. Environmental compliance and management is recognized as a multidisciplinary activity that requires large volumes of data for management and engineering decisions. Much of the data is by nature spatial or geographic. Storage, manipulation and analysis of this type of information is performed most effectively in current GIS software environments. Problems that can be addressed include facility design and siting, environmental monitoring, habitat mitigation, impact assessment and documentation, and many others. The attractiveness of adding knowledge-based expert-system software to the decision making process stems from the fact that such systems represent models of "experts" in various fields. The combination of GIS and expert-systems represents a tool that supports and enhances the reasoning and judgement process rather than only automating the procedure through prescribed computation [9], [10]. This approach will allow managers and engineers to access "expert

knowledge", covering the highly diversified set of topics associated with a pro-active environmental management strategy, easily and directly from their computer terminals.

Literature Cited

- [1] Tchobanoglous, G.H. Theiaen, and R. Eliassen. 1977. Solid Wastes, Engineering Principles and Management Issues. McGraw Hill. New York. 621 pp.
- [2] Arnoff, s. 1989. Geographic Information Systems: a Management Approach. WDL Publications. Ottawa, Ontario.
- [3] Burrough, P.A. 1986. Principles of Geographic Information Systems for Land Resource Assessment. Oxford, Clarendon.
- [4] Fedra, K. 1991. A computer Based Approach to Environmental Impact Assessment. In: Colombo, A.G. and G. Premazzi (eds). 1991. Proceedings of the Workshop on Indicators and Indices for Environmental Impact Assessment and Risk Analysis. IIASA. Austria.
- [5] Fedra, K., L. Winkelbaur, and V.R. Pantulu. 1991. Expert-systems for Environmental Screening, An Application in the Lower Mekong Basin. RR-91-19. IIASA. Austria.
- [6] Hall, C.R., C.R. Hinkle, W.M. Knott, and B.R. Summerfield. 1992. Environmental Monitoring and Research at the John F. Kennedy Space Center. J. Florida M.A. Vol 79, No 8, pp 545-552.
- [7] Stout, D.J. and R.A. Streeter. 1992. Ecological Risk Assessment: Its Role in Risk Management. The Environmental Professional. 14:197-203.
- [8] NASA. 1992. Environmental Resources Document for John F. Kennedy Space Center. KSC-DF-3080.
- [9] Summic, Z. and E. Yeh. 1992. A GIS-Based Expert-system for Residential Distribution Design. In: GIS/LIS Proceedings. San Jose, CA.
- [10] Zhu, X. and R. Healy. 1992. Towards Intelligent Decision Support: Integrating Geographical Information Systems and Expert-systems. In: GIS/LIS Proceedings. 1992. San Jose, CA.

544.43

1125 004

2527

P 1

REMOTE SENSING FOR HURRICANE ANDREW IMPACT ASSESSMENT

Bruce A. Davis
National Aeronautics and Space Administration
John C. Stennis Space Center, MS 39529

Nicholas Schmidt
Sverdrup Technology, Inc.
John C. Stennis Space Center, MS 39529

ABSTRACT

Stennis Space Center personnel flew a Learjet equipped with instrumentation designed to acquire imagery in many spectral bands into areas most damaged by Hurricane Andrew. The Calibrated Airborne Multispectral Scanner (CAMS), a NASA-developed sensor, and a Zeiss camera acquired images of these areas. The information derived from the imagery was used to assist Florida officials in assessing the devastation caused by the hurricane. The imagery provided the relief teams with an assessment of the debris covering roads and highways so cleanup plans could be prioritized. The imagery also mapped the level of damage in residential and commercial areas of southern Florida and provided maps of beaches and land cover for determination of beach loss and vegetation damage, particularly the mangrove population.

Stennis Space Center personnel demonstrated the ability to respond quickly and the value of such response in an emergency situation. The digital imagery from the CAMS can be processed, analyzed, and developed into products for field crews faster than conventional photography. The resulting information is versatile and allows for rapid updating and editing. Stennis Space Center and state officials worked diligently to compile information to complete analyses of the hurricane's impact.

REMOTE SENSING FOR URBAN PLANNING

Bruce A. Davis
National Aeronautics and Space Administration
John C. Stennis Space Center, MS 39529

Nicholas Schmidt
Sverdrup Technology, Inc.
John C. Stennis Space Center, MS 39529

John R. Jensen, Ph.D., Dave J. Cowen, Ph.D., Joanne Halls
Department of Geography
University of South Carolina
Columbia, SC 29208

Sunil Narumalani
Department of Geography
University of Nebraska
Lincoln, NE 68588

Bryan Burgess, Ph.D.
BellSouth Telecommunications Planning Department
3535 Colonnade Parkway
Birmingham, AL 35243

ABSTRACT

Utility companies are challenged to provide services to a highly dynamic customer base. With factory closures and shifts in employment becoming a routine occurrence, the utility industry must develop new techniques to maintain records and plan for expected growth. BellSouth Telecommunications, the largest of the Bell telephone companies, currently serves over 13 million residences and 2 million commercial customers. Tracking the movement of customers and scheduling the delivery of service are major tasks for BellSouth that require intensive manpower and sophisticated information management techniques. Through NASA's Commercial Remote Sensing Program Office, BellSouth is investigating the utility of remote sensing and geographic information system techniques to forecast residential development. This paper highlights the initial results of this project, which indicate a high correlation between the U.S. Bureau of Census block group statistics and statistics derived from remote sensing data.

INTRODUCTION

Utility companies, faced with increasing state and federal regulations and a growing customer base, are turning to spatial technologies to maintain records and predict future facilities expansion. The average household moves every 4.5 years, placing heavy demands on antiquated mapping and database techniques for tracking this dynamic customer base. Geographic information system (GIS) technology is becoming common throughout the utility industry as a tool to manage company resources and ensure service to its customers. Nowhere is this need more critical than with the telecommunications industry. As the primary source of communications for customer access to everything from business operations to emergency response, accurate information concerning customer requirements is critical to providing stable service. Furthermore, knowledge of changing patterns of customer location enables telephone companies to plan facility improvements so resources are in place to serve a growing population.

Nowhere is the challenge of tracking a growing customer base and planning for new service in the telephone industry more demanding than at BellSouth Telecommunications, which serves the southeastern United States. Population shifts over the past few years have resulted in dramatic increases in such metropolitan areas as Atlanta, GA and Orlando and Miami, FL, with significant increases in other areas as well. BellSouth Telecommunications currently serves over 13 million residences and 2 million business customers within a 200,000-square-mile territory in 9 southeastern states.

BellSouth maintains an extensive database on its customers and service area. However, this database lacks sophisticated information on the location of services and the spatial relationships between customers and service areas. BellSouth recognized the opportunity for using GIS and remote sensing to improve their proprietary database and in 1990 entered into a cooperative partnership project with NASA's John C. Stennis Space Center and the University of South Carolina (USC) under NASA's Earth Observations Commercial Applications Program (EOCAP). This project is a 3-year effort to develop and integrate an improved model for market forecasting. This paper will discuss a portion of the BellSouth/NASA/USC project dealing with residential land use model development and the use of remote sensing to enhance the model.

RESIDENTIAL HOUSING INVENTORY USING REMOTELY SENSED DATA

The wire center is a basic unit of geography representing the smallest division of phone service offered by the telephone utility. A wire center's geographic extent is determined by the range of cables and wires in an area served by a common three-digit telephone number prefix. Information gathered at the wire service level is aggregated to higher levels for overall analysis.

BellSouth Telecommunications needs to know the location of each single-family home, multi-family residence (duplex, triplex), apartment complex, and trailer in each wire center. These data cannot be obtained using Landsat 30 x 30 m Thematic Mapper, SPOT HRV 20 x 20 m multispectral, or even SPOT 10 x 10 m panchromatic data.^{1,2} Therefore, emphasis for this study was placed on demonstrating how BellSouth will be able to use remotely sensed image data becoming available in the later part of this decade (e.g., proposed SPOT 5 and Landsat 7 will have 5 x 5 m spatial resolution). The BellSouth model had to be developed using commercially available data so as not to base company-critical information on an unreliable source. In order to acquire data to function as a surrogate for future high spatial resolution, satellite-derived remote sensor data, NASA's Calibrated Airborne Multispectral Scanner (CAMS) was flown over the Dutch Fork Wire Center in Columbia, SC to obtain data at 5 x 5 m spatial resolution. The CAMS data were rectified to a UTM projection using high 3rd-order polynomial equations.

Various transformations of the nine channels of CAMS data were used to extract individual dwelling units from the imagery.³ The dwelling units were obtained by ratioing the data, thresholding to identify house pixels, grouping like pixels or "clumping" the house pixels, and converting the raster clumps into polygons with their own area and perimeter.

The absence (or presence) of dwelling units extracted from the 5 x 5 m CAMS data was compared with the number of units summarized in the block group statistics of the 1990 U.S. Census of Population. The results were remarkably consistent with an r^2 of .97 (Pearson product moment correlation, $r = .987$). Satellite remote sensor data having $\leq 5 \times 5$ m spatial resolution will provide valuable housing stock information for BellSouth Telecommunications when it becomes available.

Digital National Aerial Photography Program (NAPP) data are also being investigated as a possible source of housing count information.⁴ However, it is unlikely that NAPP 1:40,000 scale aerial photography obtained approximately every 5 years will meet BellSouth's yearly requirements, and currently available commercial systems do not provide sufficiently high spatial resolution imagery.

In addition to documenting existing residential dwellings, it was also important to predict where future residential development might occur. Two approaches were investigated to predict such development. First, an

empirical method was tested using selected land use planning variables and Boolean logic. Second, two predictive models based on the use of 1980 and 1990 U.S. Census of Population data, building permits, and county land use information were used to develop analytical models of future growth.

An Empirical Model of Residential Development

The following spatial variables for a portion of the Dutch Fork Wire Center were placed in a raster GIS:

- 100-Year Flood Plain Map of Lexington County
- 1992 Lexington County Land Use Zoning Map file
- 1992 Lexington County Land Use Zoning file updated using remote sensing
- Lot size computed using remote sensing data ($< \frac{1}{4}$ ac; $\frac{1}{4}$ to $\frac{1}{2}$ ac; $\geq \frac{1}{2}$ ac)
- CAMS 5 x 5 m multispectral data (for planimetric detail)

These data were then queried using Boolean logic to identify areas which have a high probability of becoming residential at specific housing densities. It is instructive to review how the variables were created and analyzed.

Few developers in South Carolina build residences in areas below the 100-year flood plain contour; therefore, these areas do not have a high probability of residential development and can be effectively removed from further analysis. County land use zoning maps are very important sources of future development information. The zoning file identifies residential development, commercial development, and those areas not yet developed. This type of information is dynamic and rapidly becomes outdated. For example, a Lexington County Zoning Map obtained for this study did not identify two major residential subdivisions already in existence. These subdivisions are easily identified on the CRT screen, and "heads-up" digitizing may be performed to update the land use zoning file.

Photogrammetrically derived length and width measurements of the residential subdivisions in conjunction with housing information (previously discussed) can be used to compute a housing density statistic per residential area. Land tracts within the wire center, which were not within the floodplain and were zoned "ready for development," were assigned to the closest housing density category in geometric space. Finally, the actual land available for residential development and its predicted density were depicted. The area of each potential tract can be computed and used to determine the average number of homes which can be located in a tract. Such predictive information is very important to BellSouth and can be obtained using relatively straightforward remote sensing and GIS technology.

An Analytical Model of Residential Development

The decennial national census provides a wealth of residential information. Typically, the longer the time from the last census, the greater the amount of error in population estimates. The fundamental need is to generate accurate inter-census inventories of urban development and to estimate where new development will occur. In order to meet these needs, an analytical study is underway to develop an integrated GIS and remote sensing environment that can be used to monitor urban expansion between census periods over large geographic areas. It addresses the need to capture systematically and analyze a wide range of data sources that are *surrogates* of urban development. A study area centered on the Columbia, SC Metropolitan Statistical Area (MSA) consisting of Lexington and Richland counties was used to test the methodology. Although the MSA contains 20 incorporated cities, 85% of the land within these counties is undeveloped.

The most important indicator of residential development is the change in the number of housing units. These data are tabulated by the Bureau of the Census every ten years. Therefore, it is possible to use the change in number of houses between 1980 and 1990 as a benchmark for examining other indicators of urban change. The 1980 and 1990 census tract polygons were chosen as the unit of measure because the two dates had similar boundaries and geographic extent. Once a geographic correspondence in tract polygons and tract

number was established, the 1990 housing counts were linked to the 1980 tract polygons. The total population change from 1980 to 1990 in Richland and Lexington Counties was +40,221 and the total housing change was +34,017. Twenty-four census tracts lost housing between 1980 and 1990, mainly in the downtown area of Columbia. The areas with the greatest increase in housing units were located in the suburbs in northeast, northwest, and west Columbia. The goal of the research was to compare surrogate measures of urban change that could be used to estimate the rate of change in number of houses and the spatial distribution of these changes. The first surrogate was based on the use of remotely sensed data.

Forecasting Residential Land Use Change Using Satellite Remote Sensor Data

Because no 1980 land cover information derived from satellite data existed, 1976 U.S. Geological Survey land use and land cover data derived from 1:60,000 aerial photography were used as the initial baseline. These data represent an inexpensive source of land cover data available for the entire United States.⁵ The polygonal land cover data are summarized in seven Level II classes.

Previous research has shown that SPOT imagery can be used in detecting urban fringe growth.^{6,7,1} A 1990 SPOT image was analyzed using traditional unsupervised classification techniques to identify the same seven land cover classes.⁸ The algorithm used to convert the 1990 raster data to polygonal data was an edge-stepping algorithm.^{9,10} To determine the amount of change from 1976 to 1990, the two classifications were intersected to create a composite land cover change file.¹¹ The composite layer was intersected with the 1980 census tract polygons to determine the land cover change by tract. The amount of change from rural to urban was obtained by performing a polygon overlay analysis. This operation identified all polygons which were undeveloped land in 1976 and developed land in 1990. These data were moderately correlated with housing change at the Census tract level ($r = 0.68$). The predictive model

$$\% \text{Change in Housing} = .005 + .538 (\% \text{Change in developed land})$$

provides a useful initial analytical model for forecasting future housing. Using this model, it is possible to estimate housing by monitoring land cover changes from remote sensing sources.

Forecasting Change Using Building Permit Applications

Another analytical approach to monitoring urban change was based on building permit transactions. The Central Midlands Regional Planning Council maintains a tabular database of all building permits issued by Richland and Lexington County, including number, street, city, county, month, year, school district, demolition, tract number, cost, type of construction, number of units, subdivision name, tax map number, and number of permits. To convert the building permit data into a GIS database, it was necessary to locate each permit geographically and link the attributes to the location. This geocoding process utilized a new BellSouth proprietary address-matching tool.

Between 1981 to 1990, 15,975 building permits were issued in the study area. Using the proprietary geocoding methodology, 67% of these permits were successfully address matched. These points were then aggregated at the census tract level and compared with the estimated housing change over the past decade. From these data, a regression procedure was used to generate a predictive analytical model for the period from 1980 to 1990:

$$\% \text{Change in Housing} = .002 + .883 (\text{change in building permits}).$$

A high correlation ($r = 0.84$) suggests that tracking building permit data provides a good way to monitor housing changes. The EOCAP II project research is now focusing on the relationship between the satellite-based residential change detection and the building permit activities.

CONCLUSION: DATA INTEGRATION

The benefits of these approaches to monitoring and modeling urban changes are evident when the various types of data are integrated into a GIS data base. While the census provides indispensable data for urban forecasting, of necessity it is aggregated both spatially and temporally. Both the remotely sensed data and the building permit information can be examined on a continuous basis at a much finer level of spatial detail than census areas. The power of this type of data integration is readily apparent when individual neighborhoods are examined. Point level data (building permits, commercial firms, and retail center locations), linear features (highways, water lines, and sewer lines), various areal features (including census polygons), and change in urban land use may be integrated. The EOCAP project is now concentrating on the creation of a robust housing model that incorporates these data sources into an improved wire center forecast.

REFERENCES

1. Jensen, J. R., E. W. Ramsey, J. M. Holmes, J. E. Michel, B. Savitsky, and B. A. Davis, 1990, "Environmental Sensitivity Index (ESI) Mapping for Oil Spills Using Remote Sensing and Geographic Information System Technology," *International Journal of Geographical Information Systems* 4(2): 181-201.
2. Haack, B. and J. R. Jensen, 1994, "Chapter 16: Urban Analysis and Planning" in the *Manual of Aerial Photointerpretation*, 2nd Ed., Falls Church: American Society for Photogrammetry & Remote Sensing, in press.
3. Cowen, D. J., J. R. Jensen, J. Halls, M. King, S. Narumalani, B. Davis, and N. Schmidt, 1993, "Estimating Housing Density with CAMS Remotely Sensed Data," *Proceedings of ACSM/ASPRS Annual Convention*, February, 1993, New Orleans, LA.
4. Light, D., 1993, "The National Aerial Photography Program as a Geographic Information System Resource," *Photogrammetric Engineering & Remote Sensing*, 59(41), 61-65.
5. U.S. Geological Survey, 1986. *Land Use and Land Cover Digital Data From 1:250,000 and 1:100,000 Scale Maps: Data Users Guide 4*. Reston: U.S. Dept. of the Interior, U.S. Geological Survey, National Cartographic Information Center.
6. Colwell, R. N., 1985, "SPOT Simulation Imagery for Urban Monitoring: a Comparison with Landsat TM and MSS Imagery and with High Altitude Color Infrared Photography," *Photogrammetric Engineering and Remote Sensing*, 51, 8, 1093-1101.
7. de Brouwer, H., C. R. Valenzuela, L. M. Valencia, and K. Sijmons, 1990, "Rapid Assessment of Urban Growth Using GIS-RS Techniques," *ITC Journal*, 1990(3): 233-235.
8. Jensen, J. R., 1986, *Introductory Digital Image Processing: A Remote Sensing Perspective*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 282 p.
9. Piwowar, J. M., E. F. LeDrew, and D. J. Dudycha, 1990, "Integration of Spatial Data in Vector and Raster Formats in a Geographic Information System Environments," *International Journal of Geographical Information Systems*, 4(4):429-444.
10. Vander Knapp, W. G. M., 1992, "The Vector to Raster Conversion: (Mis)use in Geographical Information Systems," *International Journal of Geographical Information Systems*, 6(2):159-170.
11. Jensen, J. R., D. Cowen, J. Althausen, S. Narumalani, and O. Weatherbee, 1993, "An Evaluation of CoastWatch Change Detection Protocols in South Carolina," *Photogrammetric Engineering & Remote Sensing*, (June), in press.

**REMOTE SENSING AND THE MISSISSIPPI
HIGH ACCURACY REFERENCE NETWORK**

Mark Mick
National Aeronautics and Space Administration
Commercial Remote Sensing Programs
Stennis Space Center, MS 39529

Timothy M. Alexander
Space Development Services, Inc.
Santa Fe, NM 87501-2732

Stan Woolley
Sverdrup Technology, Inc.
Stennis Space Center, MS 39529

ABSTRACT

Since 1986, the National Aeronautics and Space Administration's (NASA) Commercial Remote Sensing Program (CRSP) at Stennis Space Center has supported commercial remote sensing partnerships with industry. CRSP's mission is to maximize U.S. market exploitation of remote sensing and related space-based technologies and to develop advanced technical solutions for spatial information requirements. Observation, geolocation, and communications technologies are converging and their integration is critical to realizing the economic potential for spatial informational needs. Global Positioning System (GPS) technology enables a virtual revolution in geo-positionally accurate remote sensing of the earth. A majority of states are creating GPS-based reference networks, or High Accuracy Reference Networks (HARN). A HARN can be densified for a variety of local applications and tied to aerial or satellite observations to provide an important contribution to geographic information systems (GIS). This paper details CRSP's experience in the following areas: (1) design and implementation of a HARN in Mississippi, and (2) design and support of future applications of integrated earth observations, geolocation, and communications technology.

INTRODUCTION

The NASA CRSP at Stennis Space Center supports the spatial information industry through a data acquisition and processing infrastructure. In conjunction with its partnerships with companies and the private sector, the CRSP acquires and processes aerial and satellite imagery for use in the development of geographic information systems. These data are georeferenced, a process by which the geometry of image areas are made planimetric, and entered into a data base, where the layers of data are tied to local reference grid systems. Often the control-point intersections of several data layers do not overlay precisely (Figure 1) because the data are derived from a variety of sources. This occurrence is referred to as the misregistration of data layers. In 1991, in an effort to correct the misregistration problems encountered at Stennis, the CRSP decided to densify the local reference network around the site and its neighboring region using technology provided by the Global Positioning System.^{1,2} At this time the CRSP became aware of the National Oceanic and Atmospheric Administration (NOAA) National Geodetic Survey's (NGS) endeavor to implement the HARN on a

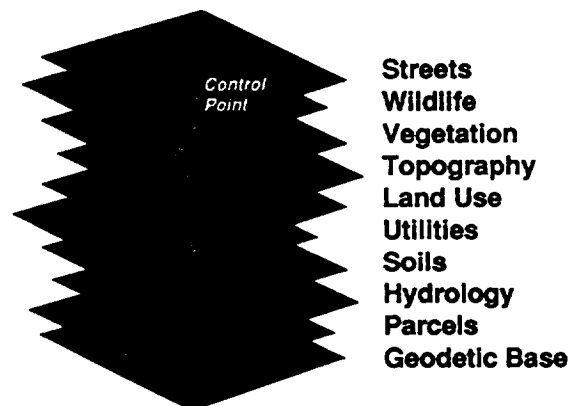


Figure 1. Misregistered Data Layers

nationwide basis. The implementation of a HARN in Mississippi was seen as the answer to Stennis' needs, in that not only would a more accurate reference network be established, but the network would be tied to a national and universal reference system. Its usage would enable misregistration errors to be minimized.

The implementation of the HARN allowed the Stennis infrastructure to capitalize on increased geolocational accuracy. This infrastructure includes an A-order station (providing locations to within a few millimeters), a more dense grid of B-order stations (providing locations to within a couple of centimeters) along the Mississippi Gulf Coast, a permanent GPS base station, a site GIS for facilities management and environmental monitoring, and an airborne sensor upgrade that provides geolocational referencing of imagery to the HARN.

Some of the monumented stations used in the MS HARN were required to be located on and in close proximity to the Stennis site. As part of the 63-station network, one A-order and one B-order monument were installed at Stennis. The A-order site is one of four A-order stations in the MS network, and the B-order station serves as the location benchmark for a dual-frequency survey quality GPS base station. This base station provides storage and archival capability for the satellite-broadcast GPS data and will support coastal area remote sensing activities and regional public and private terrestrial, airborne, and space applications of GPS technology.

Stennis is building an environmental monitoring and facilities management data base. Layers in the GIS data base include data derived from maps, satellite remote sensing, Calibrated Airborne Multispectral Scanner imagery, and continuing in-situ surveys and observations. Utilizing the HARN to georeference these layers will minimize misregistration errors during development of the GIS. The environmental portion of the data base establishes a baseline for water, soil, and air quality and for flora and fauna assemblages for the site, while the facilities management portion of the GIS contains layers detailing positions of buildings, fire hydrants, power lines, and other infrastructure components.

The NASA/Stennis aircraft, used to support a multitude of remote sensing applications, has been upgraded with a GPS unit. This configuration provides the capability to georeference imagery acquired by the Airborne Terrestrial Applications Sensor (ATLAS) to terrestrial locations referenced to the HARN. The ATLAS serves as a testbed for evaluating remote sensing products and developing specifications for airborne and spaceborne remote sensing instrumentation for dedicated applications.³ In conjunction, an optical target at the Stennis A-order station will be used for calibrating the registration of aircraft imagery for remote sensing projects.

A host of remote sensing applications will benefit from the highly accurate positional information provided by the HARN implementation. Four of these applications will be highlighted later in the paper.

HARN OVERVIEW

To correct inaccuracies in the obsolete geodetic North American Datum of 1927, in 1980 the NGS began updating the precision of latitudes, longitudes, and elevations for benchmark stations throughout the United States.⁴ When completed, the new network was called the North American Datum of 1983 (NAD83). The HARN being implemented on a state-by-state basis readjusts these coordinates and removes most of the remaining distortion from the NAD83.

The HARN is a cooperative program that utilizes the technology brought about by GPS. The NGS leads the federal initiative but increasingly relies on participants within a state to implement and augment the process. The network consists of physical reference stations, usually with 50 to 100 km spacing, whose horizontal positions relative to one another and to the NAD83 reference coordinate system are known with very high accuracy.⁵ Almost 80% of the states are either working toward or have implemented an integrated network (Figure 2). The reference stations or monuments and their associated documentation increase the efficiency of differential GPS (DGPS) correction by providing a reference for a base receiver. DGPS utilizes a base receiver (set over a station monument) that calculates the combined error in the GPS satellite range data. That correction can be applied to all other GPS receivers in the same area, eliminating the majority of error in their

measurements. The occupation time required to establish differential correction data is reduced to an absolute minimum while providing the highest degree of accuracy possible. The HARN thus becomes the standard or foundation for GPS applications within that region. The network also makes the implementation of a denser network of stations within a state easier and cheaper to establish, thus improving the efficiency of the network for a variety of users.

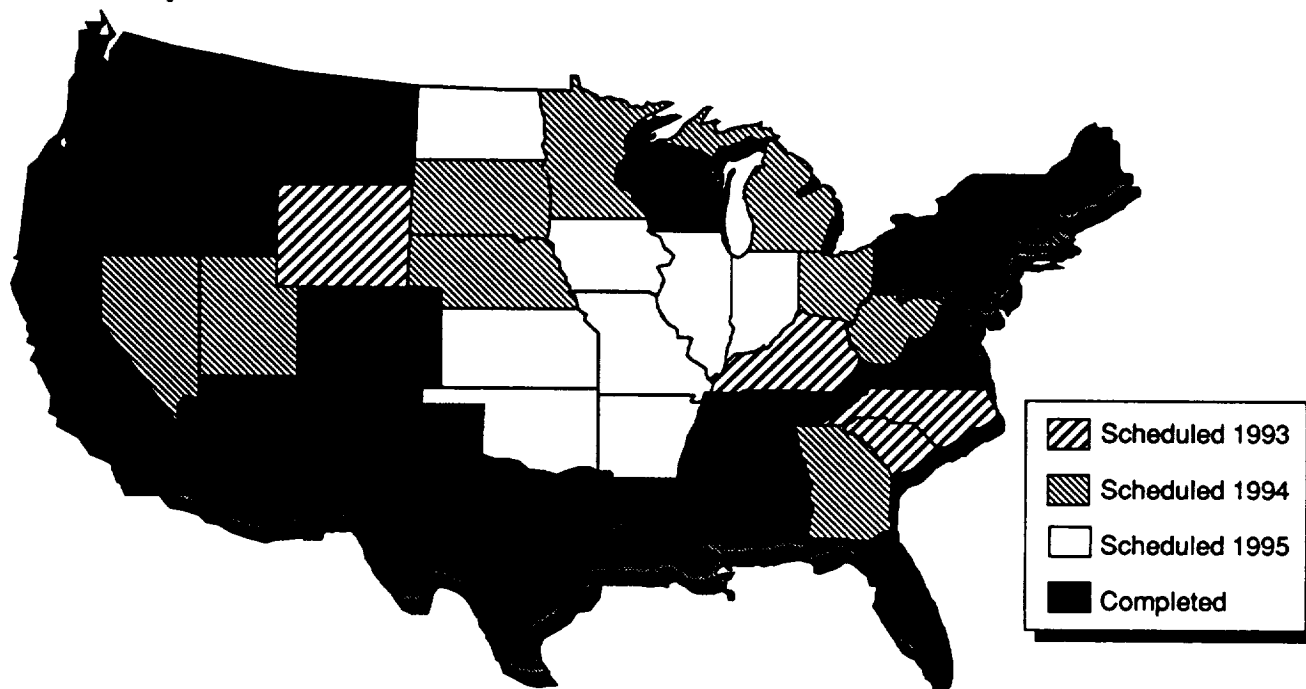


Figure 2. Statewide High Accuracy Reference Networks, September 1993

THE MISSISSIPPI HARN

The Mississippi HARN was established through a grassroots initiative that employed innovative approaches to project leadership, incentives, participation, and implementation. The effort was sponsored by NASA's Office of Advanced Concepts and Technology and represents the first time that a federal agency provided the impetus and took the initial lead to support HARN development in a state. The NASA/Stennis CRSP set the following minimum criteria for the successful establishment of this reference system:

- Priority NGS support for Mississippi,
- Support from at least one Mississippi state agency, and
- Support from one other federal agency.

Initial contacts with NGS were made in April 1992. The HARN groundwork was laid and other federal and MS state agencies were contacted. By mid-summer, the Stennis project team met at the center with representatives from NGS and other federal, state, and county agencies. Mutual responsibilities and the scope of work were established. A consensus emerged that a higher-density network would benefit a broad range of users in the state. Subject to a cooperative effort led by the State of Mississippi, a decision was made to expand the MS HARN from approximately 20 stations to 63 stations. The Mississippi Department of Transportation (MSDoT) committed itself as the lead agency for the state, and the U.S. Geological Survey (USGS) Regional Field Office in Baton Rouge, Louisiana, supported the project from the beginning, thus satisfying two of NASA's success criteria.

In late September 1992, the CRSP hosted an NGS workshop/HARN planning session at Stennis. Participants supported the project through commitments of manpower, GPS receivers, and other field equipment. A relatively simple cooperative agreement between NASA/Stennis and NOAA/NGS established the mechanism of NASA/Stennis funding to augment the development of the 63-station network. Local interest in the network was high. The original federal effort grew to a program that included participants from 21 federal, state, and local government agencies, universities, and private companies. All were eager to share in the "ownership" of the HARN (Figure 3).

Participants

- Mississippi Department of Transportation
- Mississippi Department of Environmental Quality
- Mississippi State University
- Mississippi Association of Registered Surveyors
- Pike County, Mississippi
- Harrison County, Mississippi
- Jackson County, Mississippi
- EMC, Inc.
- Electro National Corporation
- Louisiana Department of Transportation and Development
- Navigation Electronics, Inc.
- Vernon F. Meyer & Associates
- Space Development Services
- U.S. Geological Survey
- U.S. Corps of Engineers
- Naval Oceanographic Office
- U.S. Environmental Protection Agency
- National Ocean Service, Coast and Geodetic Survey
- NASA Stennis Space Center
- Johnson Controls World Services, Inc.
- Sverdrup Technology, Inc.

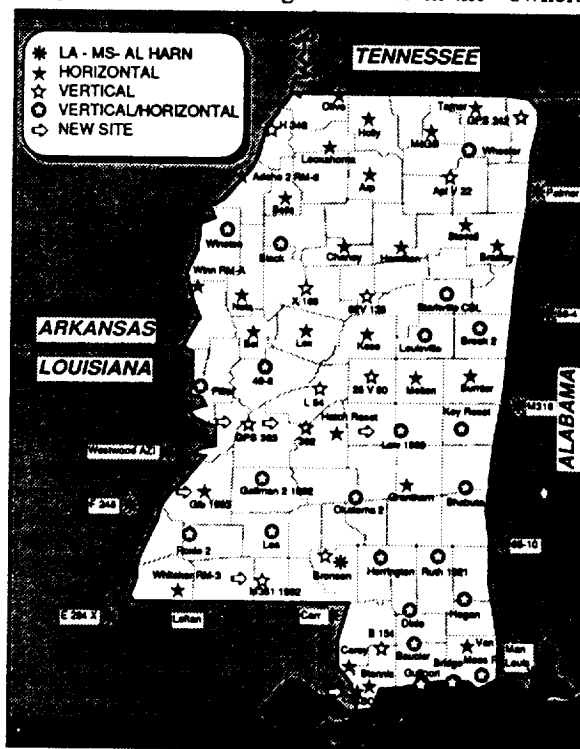


Figure 3. Mississippi HARN Stations and Participants

Sites were selected throughout the state and reconnaissance visits were made for each station by the end of November. These sites were spaced approximately 50 km apart, with more dense spacing in the Gulf Coast region and around the Stennis facility. A two-day project review and readiness session was held at the MSDoT facilities in Jackson in the middle of January 1993. Soon thereafter, confirmation of a May 3 start date for field implementation was received from NGS Headquarters. A firm schedule was established to prevent an erosion of support from participants because of other commitments.

Prior to the start date, the MSDoT hosted a three-day training session.⁶ The MSDoT provided facilities for the NGS crew center of operations and supported data reduction operations throughout the field operations.

Field observations began on May 3, 1993. Data were acquired for five and one-half hours on each of two days from those GPS satellites that were visible at each site. The observers moved to the next scheduled location the following day and acquired data. Data received during each session were downloaded and sent to Jackson, MS for a preliminary quality check.

Following this plan, all regular observations were completed in three weeks. Observations were taken for three additional days at each of the four MS A-order stations. Data were collected concurrently at A-order stations in adjacent states for regional geolocational accuracy. The observations recorded during the project were coordinated with satellite observations at selected North American stations of the Cooperative International GPS Network. Final reprocessing of the Mississippi data by NGS and its corrections to the NAD83 datum are scheduled to be completed in early December. At that time, the Mississippi HARN data (project accession number GPS-585) may be acquired from NGS at the following address:

National Geodetic Survey Information Center
SSMC3 - Code N-CG174
1315 East West Highway
Silver Springs, MD 20910-3282
Telephone: (301) 713-3242
Fax: (301) 713-4172

The HARN project in Mississippi exceeded its initial success criteria of involvement. The NGS and its State Advisors responded quickly and effectively to support the interests of Stennis and the State and helped coalesce the eventual consortium. The MSDoT took the lead of state agencies and the USGS and the Naval Oceanographic Office provided significant support as federal agencies. The workshops, open to all interested parties, acted as a catalyst for program growth. The motivation among the 21 participating agencies and organizations and their cooperative interaction provided the ingredients for an overwhelmingly successful project, moving from inception to implementation in less than 18 months. In addition, the implementation of the MS HARN was coordinated with a reference network leveling (elevation) party that travelled through the western and southern part of the state, including the Stennis site.

Throughout the term of the Mississippi project, the 21 participating agencies exhibited enthusiastic cooperation. This spirit of teamwork has created an institutional reference network among these agencies. This institutional network should facilitate continued cooperation and foster increased HARN benefits, across the state and the region, in spatial information system-related activities.

APPLICATIONS OVERVIEW

Geolocational knowledge is vital to the successful application and increased value of remote sensing technology and other geographic information. Until very recently, accurate geolocational knowledge was either quite expensive to acquire (surveyed field targets for aerial photogrammetry) or was unavailable as a practical matter (e.g., ± 40 feet point accuracy was the best available national map USGS Quadrangle quality). Satellite remote sensing data, often the only source of overseas information, is rarely registered to the earth at even course accuracies.

The Global Positioning System, knowledge of the geoid, and industry innovations in GPS equipment have revolutionized the means to acquire geolocational information. Geolocation knowledge may now be acquired from spacecraft, aircraft, terrestrial and marine reference networks, and moving land vehicles. Jack Dangermond's "instrumented universe" is becoming a reality.⁷

This revolution in the acquisition of geolocation has occurred largely within the past five years. During this very short span, technology evolution and continuous cost reductions have outpaced applications development. Widespread public and commercial applications of new geolocation knowledge capability are just beginning to emerge. This section of the paper looks ahead from building the Mississippi HARN toward a future productive era of remote sensing and related applications of the new geolocational tools afforded by GPS. This brief look ahead is organized under four applications support areas that are either planned or underway at the Stennis Space Center.

Commercial Applications and Systems Testbed

Six years of remote sensing partnerships between NASA/Stennis and over fifty different companies underscore the importance of several testbed-related factors to successful commercial innovation. Successful commercial remote sensing innovation takes into account technical, financial, and other considerations of the entire acquisition-to-delivery system. Most individual remote sensing products, processes (e.g., software modules), and services are affected by dependencies (or competitive threats) involving data acquisition, data handling and distribution, data processing and integration, and options for customer packaging and delivery.

A good example of the importance of system-wide knowledge to the development of new products or services may be drawn from GPS applications in highway inventory. Major investments have been made in perfecting van-mounted highway inventory systems.⁸ Early commercial prototypes using Coarse Acquisition Code GPS aided by gyros produced road alignment and directional information as the key products. These products possessed accuracies of about 15 meters and a Quadrangle View of the highway applications market. Before these applications had matured, they were superseded by DGPS applications (<5 meter accuracies) that required another system component - telecommunications. DGPS applications were in their infancy when they, in turn, were augmented by video frame camera systems.

Now that GPS is being successfully integrated with airborne scanning and photogrammetry, a fundamental reassessment of the roadway inventory market seems in order. The initial impetus for van-mounted GPS inventory may have been succeeded by advances in other areas of remote sensing.

Product or service innovation demands understanding the offering's entire systems context. Awareness and/or use of an end-to-end spatial information testbed is one key source for understanding product systems. The Airborne Instrument Test System is an end-to-end data acquisition and processing system at Stennis (Figure 4). The system provides all the engineering and computational support required for scientific and commercial remote sensing development.⁹ The HARN and associated spatial data infrastructure allow the CRSP to quantify and compare alternative technology approaches to market solutions.

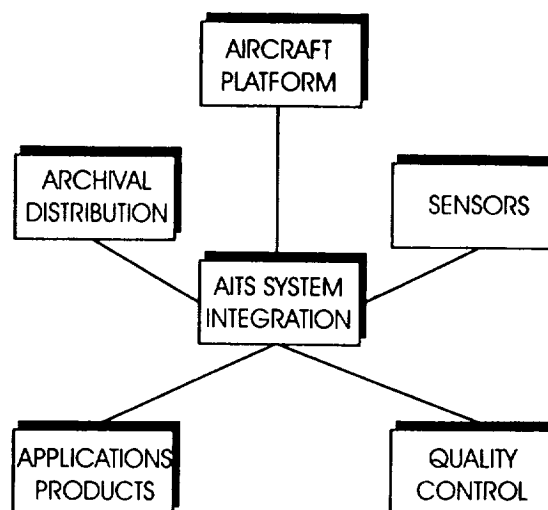


Figure 4. Airborne Instrumentation Test System Components

Successful improvements in new products, processes, or services must be demonstrated or perfected in a realistic rather than an ideal applications environment. Demonstrations and adaptations ideally take place with the prospective customer in the application setting. Prior to a customer's exposure, however, many prospective offerors would rather develop proprietary testbed demonstrations of their product, process, or service. Such tests of products are especially important when they are targeted toward completely new markets where an initial visible "failure" may close the door to customer interest. The early claims of airborne scanning, which failed to take into account the complexities of post-acquisition image processing, are good examples of the need to prototype the entire chain of dependencies from acquisition to product delivery.

Most companies and even government agencies lack the means for controlled test and evaluation of new products. The CRSP testbed has been used to prototype equipment, software, and subsystems. The HARN and its associated spatial data infrastructure will enable the CRSP to support local and regional prototype tests.

Public Applications in Pike County, Mississippi

The CRSP has found that progressive government agencies are sometimes the first entities willing to experiment with emerging remote sensing technologies. Numerous position inconsistencies in the San Diego County, CA infrastructure GIS prompted county officials to support the development of a more accurate regional ground control network in April 1991.¹⁰ In contrast, Mississippi's rural counties often do not have the same level of infrastructure support as do more populated counties such as San Diego. However, Pike and Hancock counties in Mississippi are two exceptions to this rule. Both participated in the design and implementation of the

Mississippi HARN. (For purposes of brevity, only Pike County's network application initiative is discussed here.)

The Pike County Tax Assessor's Office has lead the GIS activities for the county. This office has experimented with GPS-based roadway applications since 1988 and has provided its own testbed for van-mounted GPS applications (Figure 5). Magnolia, a small city in Pike County, has been the setting for extensive practical tests of GPS-based roadway alignment and inventory.¹¹ The Assessor's Office has developed and operates its own GIS with little outside assistance. This GIS has been bootstrapped at a low cost from the ground up with only minimal support from the county's infrastructure.

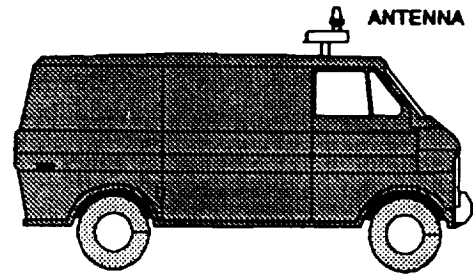


Figure 5. Van Testbed

Pike County has several applications which require the use of GPS technology. The County is currently constructing infrastructure layers for its GIS data base. Precise geolocation of such features as roads, right-of-ways, and service networks is essential for the implementation of their GIS, particularly in the few urban centers. Pike County intends to utilize this data base to determine suitable locations for industries it would like to attract to the area. Information regarding the costs of connecting to services, improving access, and acquiring properties may be readily determined using a GIS only if information regarding existing infrastructure is accurate. GPS survey techniques using the HARN provide the necessary accuracy and co-registration for various infrastructure layers.

The major portion of Pike County's tax base is generated through rural property assessment and is used largely for agriculture and silviculture (forest management). Property assessments are based on actively utilized acreage of farm lands. Such assessments require knowledge of property boundaries and activities currently in progress on the lands. The former can be obtained from existing parcel maps with GPS surveys to ensure consistency and accuracy. The latter can be obtained only through the collection and analysis of imagery. Aerial photogrammetric surveys acquired and archived by the county are over a decade old. Pike County hopes to incorporate new aerial imagery to develop a land use classification. On completion of this new base map, the county intends to investigate the utility of satellite data to develop a cost-effective means of assessing changes in land use practices on an interim basis. Such information can be used in conjunction with directed ground GPS surveys to update the imagery-based map in a cost-effective fashion.

A successful combination of vehicle and airborne spatial data acquisition referenced to the HARN and its incorporation into a GIS tax base has widespread applicability in Mississippi and elsewhere. The planned applications in Pike County may provide valuable verification for commercial GPS services and a practical test of the value of CRSP's new ATLAS scanner.

Public agency innovation in the use of the HARN, photo and digital imagery, and GIS is one key source of convergent technology applications. Pike County, through its affiliations with industry and NASA, demonstrates that significant convergent technology innovations are possible at a low cost.

Applications to State Highway Transportation

The Mississippi DoT was the lead state agency in the design and implementation of the Mississippi HARN. The MSDoT lead role in project development was more a matter of foresight than a response to proven requirements because the Department had little internal experience with GPS prior to the HARN initiative. The MSDoT benefitted from close working relationships with its NGS State Advisor and the experiences of the Louisiana and Alabama DoTs, both of which were completing their HARN development process in 1992.

Now that the MSDoT has been instrumental in making the HARN a reality in the state, what are the applications paths it is able to take? The first path will be continued training of field crews and Headquarters

staff in the use of the GPS receivers and personal computer processing equipment. Traditional applications involve right of way, roadway expansion or realignment, new entrances, and other pre-construction activities.

Training and experience in the productive use of the HARN will be critical to realizing the MSDoT's benefits from a network. One key to quick realization of benefits is to apply GPS to traditional DoT survey requirements. Reports from the Louisiana Department of Transportation and Development indicate that the use of HARN and GPS in that state led to cost savings through the eradication of survey duplication in road alignment surveys.

The second path will be strengthened leadership in statewide GIS coordination and development. The MSDoT will lead the state's GIS coordination group in 1993-94 from a position of having led a successful multi-agency HARN development program. In this sense, the network helped create a successful model of institutional cooperation and investment in spatial data infrastructure.

The HARN process itself placed its lead state agency in a stronger position to coordinate and lead a statewide GIS strategy. From a technology applications standpoint, the major issue to be resolved in Mississippi will be adoption of the network as the spatial data reference standard for the coming decade.

The third path involves future innovation built on HARN and GPS applications to traditional MSDoT requirements. The tangible success of moving the network from concept to reality in less than two years provides support for continuing innovation in the state and within the MSDoT. The specific paths for future innovation and co-investment are less clear than the willingness to tackle the next challenge. What seems clear at the concept level is that the MSDoT will move from survey to inventory applications of GPS and HARN. From a technology evolution standpoint, a convergence of observation and geolocation technologies enables the MSDoT and other state agencies to consider realistically a statewide GIS data base for environmental and developmental decisions.

The HARN process and its successful outcome not only provide a basis for achieving short-term benefits for the MSDoT, but may enable the MSDoT to lead a broader application of convergent technology to the development and environmental concerns of Mississippi.

Bringing Satellite Remote Sensing "Down to Earth"

These are exciting times for commercial satellite remote sensing. With the close of the Cold War, the Federal Government appears willing to consider the licensing of National Security satellite technology for civil-commercial use. At the same time, the Land Remote Sensing Commercialization Act was amended in 1992 to allow different treatment for truly commercial systems. These two government policies coincide with significant growth in GIS markets for remote sensing data worldwide and a proliferation of products that can take advantage of these policy shifts.

As 1993 comes to a close, one application for a commercial remote sensing satellite system has been granted and another has progressed toward approval. Either or both of the World View and Lockheed systems would dramatically change the character of space-derived data and the applications of such data to public or commercial uses. World View proposes three-meter spatial resolution and Lockheed's Commercial Remote Sensing System proposes stereo spatial resolution at the one-meter level.

Satellite remote sensing (Figure 6) will truly be brought down to earth if either or both of the proposed systems are approved and launched. Geolocation technology will play a crucial role in realizing the potential of these new data sources for GIS applications. GPS and star sensors on board the satellites enable a design goal of one pixel registration to be pursued, where one pixel is a picture element having both spatial and spectral properties. (The spatial variable defines the apparent size of the resolution cell and the spectral variable defines the intensity of the spectral response for that cell.) The HARN system in the U.S., and the growth of portable DGPS system will allow sub-pixel accuracy reference to be built quickly and at low cost. On a global scale, GPS and its reference systems provide not only the basis for accurate geolocation registration of spaceborne data, but a

universal means to attach ground truth to orbital or airborne observations.

CRSP plans to use its entire spatial data infrastructure to help verify and develop the opportunities provided by changes in U.S. policy coupled with commercial technology and market incentives. The applications which appear most beneficial to U.S. companies are high resolution spatial and spectral test data sets, experimentation with HARN and related forms of visual registration of digital imagery, and customer/market tests of data and information that simulate future space-derived products.

The applications of existing HARNs, the development of visual HARNs, airborne and terrestrial GPS, and product simulations will be critical to moving satellite imagery from the domain of the scientist and specialist to the domain of the customer. With increasing demand from its industrial partners, CRSP is planning up to ten projects to demonstrate the utility of the HARN to remote sensing applications.

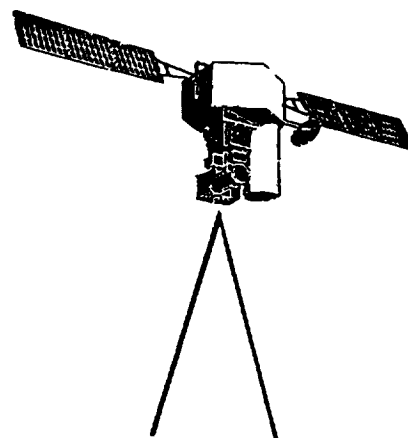


Figure 6. Satellite Remote Sensing

SUMMARY

In addition to the benefits provided to Stennis organizations for applications research, development, and operational purposes, outside GPS users will be able to utilize the HARN to increase the accuracy and timeliness of their products. Once the NGS publishes the final computations for each Mississippi site (scheduled for November 1993), the HARN will be used by the CRSP to demonstrate the viability of a highly accurate reference configuration for GIS base maps. GIS data layers can be registered with universal accuracy, allowing adjacent areas to form a seamless mosaic and avoiding controversy over discrepancies in boundaries and site perimeters. The Mississippi HARN will provide NASA, the state of Mississippi, and private industry with the basic horizontal reference controls necessary to support a host of remote sensing/spatial information system applications, benefitting all parties concerned.

ACKNOWLEDGEMENTS

This project was performed under the Commercial Remote Sensing Program at Stennis Space Center for the NASA Office of Advanced Concepts and Technology. The work performed by Sverdrup Technology, Inc. was under NASA Contract No. NAS13-290, Technical Work Request VXD09315. Mention of particular products, companies, or agencies does not imply endorsement by the U.S. government, NASA, or Sverdrup Technology, Inc. This information is provided for the benefit of the reader. The authors wish to thank the following individuals and their staffs for their contributions to the success of the Mississippi HARN implementation: John Love, Gilbert Mitchell, Pam Fromhertz, Roy Anderson, and David Doyle of NGS Headquarters; Don Rexrode, Bob Zurfluh, and Bill Rindal, NGS State Advisors to Mississippi, Louisiana, and Alabama, respectively; Chuck Hill of Stennis' CRSP; Marlin Collier and Randall Fulcher of the MSDOT; Frank Kenney of the USGS; Steve Oivanki of the MS DEQ; John Knight and Tom Berry of the Army Corps of Engineers; Roffie Burt of Mississippi State University; Allan Lemley and Marc Chalona of the Naval Oceanographic Office; Mark Mattox of EMC, Inc.; and all Mississippi state and county personnel that supported the field surveys. Much appreciation goes to Charlie Poche and Joe Strickland of Navigation Electronics who donated three Trimble 4000 SSE units for use in the field activities.

REFERENCES

1. Rocken, Christian, and Thomas M. Kelecy. 1992. "High-Accuracy GPS Marine Positioning for Scientific Applications." *GPS World*, June, pp. 42-47.
2. Gilbert, Chuck. 1993. "Portable GPS Systems for Mapping: Features Versus Benefits." *Earth Observation Magazine*, October, pp. 43-48.
3. Birk, Ronald J., and Bruce Spiering. 1992. "Commercial Applications Multispectral Sensor System." *Small Satellite Technologies & Applications II*, SPIE Vol. 1691, 49-59.
4. Olsen, Norman T. 1993. "Update Your Database - The North American Continent Has Moved." *GIS World*, September, pp. 40-41.
5. Strange, William E., and John D. Love. 1991. "High Accuracy Reference Networks - A National Perspective." Presented at ASCE Specialty Conference - Transportation Applications of GPS Positioning Strategy, Sacramento, CA, September 18-21.
6. Grunthal, Captain Melvyn C. 1993. *INSTRUCTIONS: Mississippi HARN, 1993*.
7. Dangermond, Jack. 1992. Presentation to the First GIS-in-Business Conference, Denver, CO.
8. Center for Mapping, Ohio State University. 1991. *GPS/Imaging/GIS Project*. Columbus, OH.
9. Hill, C. L., R. J. Birk, E. Christensen, and T. Alexander. 1991. "Airborne Instrument Test System (AITS) Program." *Proceedings of American Society for Photogrammetry and Remote Sensing*.
10. Petersen, Carolyn. 1992. "Precisely San Diego." *GPS World*, April, pp. 25-29.
11. Lewis, Robert. 1992. NAVSTAR Mapping Corporation, personal communication.

A VISUAL DETECTION MODEL FOR DCT COEFFICIENT QUANTIZATION

Albert J. Ahumada, Jr.
 Andrew B. Watson
 NASA Ames Research Center
 Moffett Field, California 94035-1000
 ahumada@vision.arc.nasa.gov
 beau@vision.arc.nasa.gov

ABSTRACT

The discrete cosine transform (DCT) is widely used in image compression, and is part of the JPEG and MPEG compression standards. The degree of compression, and the amount of distortion in the decompressed image are controlled by the quantization of the transform coefficients. The standards do not specify how the DCT coefficients should be quantized. Our approach is to set the quantization level for each coefficient so that the quantization error is near the threshold of visibility. Here we combine results from our previous work to form our current best detection model for DCT coefficient quantization noise. This model predicts sensitivity as a function of display parameters, enabling quantization matrices to be designed for display situations varying in luminance, veiling light, and spatial frequency related conditions (pixel size, viewing distance, and aspect ratio). It also allows arbitrary color space directions for the representation of color. In a further development, we have developed a model-based method of optimizing the quantization matrix for an individual image. The model described above provides visual thresholds for each DCT frequency. These thresholds are adjusted within each block for visual light adaptation and contrast masking. For a given quantization matrix, the DCT quantization errors are scaled by the adjusted thresholds to yield perceptual errors. These errors are pooled non-linearly over the image to yield total perceptual error. With this model we may estimate the quantization matrix for a particular image that yields minimum bit rate for a given total perceptual error, or minimum perceptual error for a given bit rate. Custom matrices for a number of images show clear improvement over image-independent matrices. Custom matrices are compatible with the JPEG standard, which requires transmission of the quantization matrix.

1. INTRODUCTION

1.1 DCT image compression

The discrete cosine transform (DCT) has become an image compression standard (ref. 1, 2, 3). Typically the image is divided into 8x8-pixel blocks, which are each transformed into 64 DCT coefficients. The DCT transform coefficients $I_{m,n}$, of an $N \times N$ block of image pixels $i_{j,k}$, are given by

$$I_{m,n} = \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} i_{j,k} c_{j,m} c_{k,n} \quad m,n = 0 \dots N-1 \quad (1a)$$

where

$$c_{j,m} = \alpha_m \cos\left(\frac{\pi m}{2N} [2j+1]\right), \quad (1b)$$

and

$$\begin{aligned} \alpha_m &= \sqrt{1/N} & m=0 \\ &= \sqrt{2/N} & m>0 \end{aligned} \quad (1c)$$

The block of image pixels is reconstructed by the inverse transform:

$$i_{j,k} = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} I_{m,n} c_{j,m} c_{k,n} \quad j, k = 0 \dots N-1 \quad (2)$$

which for this normalization is the same as the forward transform. Quantization of the DCT coefficients achieves image compression, but it also generates distortion in the decompressed image. If a single coefficient is quantized and its block is reconstructed, the difference between the original image block and the reconstructed block is the error image. This error image has the form of the associated basis function, and its amplitude is proportional to the quantization error of the coefficient. Since the inverse transform is linear, the error image resulting from quantizing multiple coefficients is a sum of such images.

1.2 The Quantization Matrix

The JPEG compression standard (ref. 1, 2) requires that uniform quantization be used for the DCT coefficients, but the quantizer step size to be used for each coefficient is left to the user. The step size used for coefficient $I_{m,n}$ is denoted by $Q_{m,n}$. A coefficient is quantized by the operation

$$S_{m,n} = \text{Round}[I_{m,n}/Q_{m,n}] \quad (3a)$$

and restored (with the quantization error) by

$$\hat{I}_{m,n} = S_{m,n} Q_{m,n}. \quad (3b)$$

Two example quantization matrices can be found in the JPEG standard (ref. 2). These matrices appear in Table 1 following the references. These matrices were designed for a particular viewing situation. No suggestions were provided for how they should be changed to accommodate different viewing conditions, or for compression in a different color space. Our research was initiated to provide quantization matrices suitable for compression in the RGB color representation (ref. 4). Subsequently, a theoretical framework was constructed and additional measurements have been done (ref. 5, 6, 7, 8). Here we summarize the quantization matrix design technique. It can be applied under a wide variety of conditions: different display luminances, veiling luminances, spatial frequencies, and color spaces. The basic idea of the technique is to develop a detection model that predicts the detectability of the artifacts in a perceptual space representation. This step is described in Section 2. A quantizer step size is then determined from the sensitivity of the perceptual space representation to the quantization distortion. This step is described in Section 3.

2. THE DETECTION MODEL

2.1 The Luminance Detection Model

The luminance detection model predicts the threshold of detection of the luminance error image generated by quantization of a single DCT coefficient $I_{m,n}$. We use the subscript Y for luminance, since we assume that it is defined by the 1931 CIE standard (ref. 9). This error image is assumed to be below the threshold of visibility if its zero-to-peak luminance is less than a threshold $T_{m,n}$ given by

$$\begin{aligned}
\log T_{Y,m,n} &= P(f_{m,n}; b_Y, k_Y, f_Y) \\
&= \log b_Y && \text{if } f_{m,n} \leq f_Y \\
&= \log b_Y + k_Y (\log f_{m,n} - \log f_Y)^2 && \text{if } f_{m,n} > f_Y
\end{aligned} \tag{4}$$

This function P represents a low-pass contrast sensitivity function of spatial frequency. Although luminance contrast sensitivity is more correctly modeled as band-pass, we choose a low-pass function for this application. This ensures that no new artifacts become visible as viewing distance increases. A low-pass function is also convenient because purely chromatic channels are low-pass in this spatial frequency range. We will use P for the luminance and chrominance channels of our model.

The spatial frequency, $f_{m,n}$, associated with the m,n th basis function, is given by

$$f_{m,n} = \frac{1}{2N} \sqrt{(m/W_x)^2 + (n/W_y)^2} \tag{5}$$

where W_x and W_y are the horizontal and vertical pixel spacings in degrees of visual angle. The term b_Y has three components:

$$b_Y = \frac{s T_Y}{\theta_{m,n}}. \tag{6}$$

The parameter s is a fraction, $0 \leq s \leq 1$, to account for spatial summation of quantization errors over blocks. We set it to unity to model detection experiments with only one block (ref. 5). Our summation results suggest that it should be equal to the inverse of the fourth root of the number of blocks contributing to detection (ref. 6). We suggest the value $s=0.25$, corresponding to 16×16 blocks. The factor T_Y gives the dependence of the threshold on the image average luminance \bar{Y} .

$$\begin{aligned}
T_Y &= \frac{\bar{Y}^{a_T} Y_T^{1-a_T}}{S_0} && \bar{Y} \leq Y_T \\
&= \frac{\bar{Y}}{S_0} && \bar{Y} > Y_T
\end{aligned} \tag{7}$$

where suggested parameter values are $Y_T=15 \text{ cd/m}^2$, $S_0=40$, and $a_T=0.65$.

The product of a cosine in the x with a cosine in the y direction can be expressed as the sum of two cosines of the same radial spatial frequency but differing in orientation. The factor

$$\theta_{m,n} = r + (1-r) \left(1 - \left[\frac{2 f_{m,0} f_{0,n}}{f_{m,n}^2} \right]^2 \right) \tag{8}$$

accounts for the imperfect summation of two such frequency components as a function of the angle between them. Based on the fourth power summation rule for the two components when they are orthogonal, r is set to 0.6. An additional oblique effect can be included by decreasing the value of r .

The parameters f_Y and k_Y determine the shape of P and depend on the average luminance \bar{Y} .

$$\begin{aligned} f_Y &= f_0 \bar{Y}^{a_f} Y_f^{-a_f}, & \bar{Y} \leq Y_f \\ &= f_0, & \bar{Y} > Y_f \end{aligned} \quad (9)$$

and

$$\begin{aligned} k_Y &= k_0 \bar{Y}^{a_k} Y_k^{-a_k} & \bar{Y} \leq Y_k \\ &= k_0 & \bar{Y} > Y_k \end{aligned} \quad (10)$$

where

$$f_0 = 6.8 \text{ cycles / deg}$$

$$a_f = 0.182$$

$$Y_f = 300 \text{ cd / m}^2$$

$$k_0 = 2$$

$$a_k = 0.0706, \text{ and}$$

$$Y_k = 300 \text{ cd / m}^2.$$

2.2 The Chrominance Detection Model

We now add two chromatic channels to the luminance-only model. From the large number of color spaces that have been proposed for chromatic discriminations, we have selected one close to that suggested by Boynton (ref. 9): a red-green opponent channel and a blue channel. Our channels are defined in terms of the CIE 1931 XYZ color space. The blue channel is just Z, and the opponent red-green channel O is given by

$$O = 0.47 X - 0.37 Y - 0.10 Z. \quad (11)$$

This opponent channel is approximately the Boynton (ref. 9) (Red-cone)-2(Green-cone) channel. Our model now needs the threshold for quantization noise in the O and Z channels. For simplicity, we model the chromatic thresholds by

$$\log T_{O,m,n} = P(f_{m,n}; \frac{0.36 s T_Y}{\theta_{m,n}}, k_Y, \frac{f_Y}{4}), \quad (12)$$

and

$$\log T_{Z,m,n} = P(f_{m,n}; \frac{3 s T_Y}{\theta_{mn}}, k_Y, \frac{f_Y}{4}), \quad (13)$$

These shapes of these are in agreement with experimental results of Mullen (ref. 10), except that the slopes for the chromatic channels are found to be steeper than that of the luminance channel. The reason for keeping them the same is to prevent strong quantization of purely chromatic channels, since there is a fair amount of individual variability in the exact direction of isoluminance.

Although we previously used \bar{Z} to set the level of the Z threshold (ref. 7), we are using \bar{Y} here under the assumption that the average image color is close to white and hence that they are roughly equal.

Finally, we say that the errors from the quantization of a coefficient are visible, if the error in any of the three channels is visible.

3. QUANTIZATION MATRIX DESIGN

Suppose that one color dimension D in a color space linearly related to our YOZ color space is to be quantized. Let D_Y , D_O , and D_Z be the amplitudes of the errors in YOZ space generated by a unit error in D . An error generated in the D image by quantizing the m,n th DCT coefficient is then below threshold if it is less than

$$T_{D,m,n} = \min\left(\frac{T_{Y,m,n}}{D_Y}, \frac{T_{O,m,n}}{D_O}, \frac{T_{Z,m,n}}{D_Z}\right). \quad (14)$$

The D quantization matrix entries are obtained by dividing the thresholds above by the DCT normalization constants (α_m in Equation (1c)):

$$Q_{D,m,n} = \frac{2T_{D,m,n}}{\alpha_m \alpha_n}, \quad (15)$$

The factor 2 results from the maximum quantization error being half the quantizer step size.

3.1 Quantization in YC_rCb Space

In an attempt to put all luminance information in a single channel, color images are often represented in the YC_rCb color space for image compression. (ref. 2) give the transformation from RGB to YC_rCb as

$$\begin{aligned} Y &= 0.3R + 0.6G + 0.1B, \\ C_r &= (R - Y) / 1.6 + 0.5, \\ C_b &= (B - Y) / 2 + 0.5. \end{aligned} \quad (16)$$

Suppose that the viewing conditions are set so that the average image luminance is 40 cd/m^2 , the pixel spacings are 0.028 deg , and the monitor calibration of the XYZ outputs for unit RGB inputs are given by the matrix

	X	Y	Z	
R	26.1	13.3	2.3	
G	25.2	48.9	10.2	
B	9.3	4.7	35.7	

(17)

The values of D_Y , D_O , and D_Z for each dimension turn out to be:

	D_Y	D_O	D_Z	
Y	66.9	-1.1	48.2	
C_r	-17.8	17.1	-4.5	
C_b	-7.0	0.6	67.9	

(18)

The quantization matrices appear in Table 2 following the references.

4. SUMMARY

We have presented a model for predicting visibility thresholds for DCT coefficient quantization error, from which quantization matrices for use in DCT-based compression can be designed. We regard this as preliminary results of work in progress. The quantization matrices computed by the techniques described above take no account of image content. We now show how an extension of this model may be used to optimize quantization matrices for individual images or a class of images.

5. LIMITATIONS OF THE IMAGE-INDEPENDENT APPROACH

The preceding approach constructs a quantization matrix independent of the image. While a great advance over the *ad hoc* matrices that preceded it, the image-independent approach has several shortcomings. The fundamental drawback is that visual thresholds for artifacts are dependent on the image upon which they are superimposed.

First, visual thresholds increase with background luminance., and variations in local mean luminance within the image will in fact produce substantial variations in DCT threshold. We call this *luminance masking*. Second, threshold for a visual pattern is typically reduced in the presence of other patterns, particularly those of similar spatial frequency and orientation, a phenomenon usually called *contrast masking*. This means that threshold error in a particular DCT coefficient in a particular block of the image will be a function of the value of that coefficient in the original image. Third, the image-independent approach ensures that any single error is below threshold. But in a typical image there are many errors, of varying magnitudes. The visibility of this error ensemble is not generally equal to the visibility of the largest error, but reflects a pooling of errors, over both frequencies and blocks of the image. We call this *error pooling*. Fourth, when all errors are kept below a perceptual threshold a certain bit rate will result. The image-independent method gives no guidance on what to do when a lower bit rate is desired. The *ad hoc* "quality factors" employed in some JPEG implementations, which usually do no more than multiply the quantization matrix by a scalar, will allow an arbitrary bit rate, but do not guarantee (or even suggest) optimum quality at that bit rate. We call this the problem of *selectable quality*.

6. IMAGE-DEPENDENT APPROACH

Here we present a general method of designing a custom quantization matrix tailored to a particular image. This *image-dependent* method incorporates solutions to each of the problems described above. The strategy is to develop a very simple model of perceptual error, based upon DCT coefficients, and to iteratively estimate the quantization matrix which yields a designated perceptual error or bit-rate. We call this the *DCTune* algorithm, because it tunes the DCT quantization matrix to the individual image(ref. 11, 12).

6.1. JPEG DCT Quantization

In the JPEG image compression standard, the image is first divided into blocks of size {8,8}. Each block is transformed into its DCT, which we write $I_{m,n,b}$, where m,n indexes the DCT frequency (or basis function), and b indexes a block of the image. Each block is then quantized by dividing it, coefficient by coefficient, by a quantization matrix (QM) $Q_{m,n}$, and rounding to the nearest integer

$$S_{m,n,b} = \text{Round}\left[I_{m,n,b}/Q_{m,n}\right] \quad (19)$$

The quantization error in the DCT domain is then

$$E_{m,n,b} = I_{m,n,b} - S_{m,n,b} Q_{m,n} \quad (20)$$

6.2 Luminance Masking

Detection threshold for a luminance pattern typically depends upon the mean luminance of the local image region: the brighter the background, the higher the luminance threshold (ref. 13, 14). This is usually called "light adaptation," but here we call it "luminance masking" to emphasize the similarity to contrast masking, discussed in the next section. We can compute a luminance-masked threshold matrix for each block in either of two ways. The first is to make use of a formula such as that supplied by Peterson *et al.* (ref. 7)

$$T_{m,n,b} = \text{apw}[m,n, \bar{Y} I_{00b} / \bar{I}_{00}] \quad (21)$$

where I_{00b} is the DC coefficient of the DCT for block b , \bar{Y} is the mean luminance of the display, and \bar{I}_{00} is the DC coefficient corresponding to \bar{Y} (1024 for an 8 bit image).

A second, simpler solution is to approximate the dependence of $T_{m,n,b}$ upon I_{00b} with a power function:

$$T_{m,n,b} = T_{m,n} (I_{00b} / \bar{I}_{00})^{a_T} \quad (22)$$

The initial calculation of $T_{m,n}$ should be made assuming a display luminance of \bar{Y} . The parameter a_T takes its name from the corresponding parameter in the formula of Ahumada and Peterson, wherein they suggest a value of 0.65. Note that luminance masking may be suppressed by setting $a_T=0$. More generally, a_T controls the degree to which this masking occurs. Note also that the power function makes it easy to incorporate a non-unity display Gamma, by multiplying a_T by the Gamma exponent.

6.3 Contrast Masking

Contrast masking refers to the reduction in the visibility of one image component by the presence of another. Here we consider only masking within a block and a particular DCT coefficient. We employ a model of visual masking that has been widely used in vision models, (ref. 15, 16). Given a DCT coefficient $I_{m,n,b}$ and a corresponding absolute threshold $T_{m,n,b}$ our masking rule states that the masked threshold $M_{m,n,b}$ will be

$$M_{m,n,b} = T_{m,n,b} \max \left[1, \left| I_{m,n,b} / T_{m,n,b} \right|^{w_{m,n}} \right] \quad (23)$$

where $w_{m,n}$ is an exponent that lies between 0 and 1. Because the exponent may differ for each frequency, we allow a matrix of exponents equal in size to the DCT. Note that when $w_{m,n}=0$, no masking occurs, and the threshold is constant at $T_{m,n,b}$. When $w_{m,n}=1$, we have what is usually called "Weber Law" behavior, and threshold is constant in log or percentage terms (for $I_{m,n,b} > T_{m,n,b}$). Because the effect of the DC coefficient upon thresholds has already been expressed by luminance masking, we specifically exclude the DC term from the contrast masking, by setting the value of $w_{00} = 0$.

6.4 Perceptual Error and Just-Noticeable-Differences

In vision science, we often express the magnitude of a signal in multiples of the threshold for that signal. These threshold units are often called "just-noticeable differences," or *jnd*'s. Having computed a masked threshold $M_{m,n,b}$, the error DCT may therefore be expressed in *jnd*'s as

$$D_{m,n,b} = E_{m,n,b} / M_{m,n,b} \quad (24)$$

6.5 Spatial Error Pooling

To pool the errors in the jnd DCT we employ another standard feature of current vision models: the so-called Minkowski metric. It often arises from an attempt to combine the separate probabilities that individual errors will be seen, in the scheme known as "probability summation" (ref. 17). We pool the jnds for a particular frequency m,n over all blocks b as

$$\Psi_{m,n} = \left(\sum_b |D_{m,n,b}|^{\beta_s} \right)^{1/\beta_s} \quad (25)$$

In psychophysical experiments that examine summation over space a β_s of about 4 has been observed (ref. 17). The exponent β_s is given here as a scalar, but may be made a matrix equal in size to the QM to allow differing pooling behavior for different DCT frequencies. This matrix $\Psi_{m,n}$ is now a simple measure of the visibility of artifacts within each of the frequency bands defined by the DCT basis functions. We call it the "perceptual error matrix."

6.6 Frequency Error Pooling

This perceptual error matrix $\Psi_{m,n}$ may itself be of value in revealing the frequencies that result in the greatest pooled error for a particular image and quantization matrix. But to optimize the matrix we would like a single-valued perceptual error metric. We obtain this by combining the elements in the perceptual error matrix, using a Minkowski metric with a possibly different exponent, β_f

$$\Psi = \left(\sum_{m,n} \Psi_{m,n}^{\beta_f} \right)^{1/\beta_f} \quad (26)$$

It is now straightforward, at least conceptually, to optimize the quantization matrix to obtain minimum bit-rate for a given Ψ , or minimum Ψ for a given bit rate. In practice, however, a solution may be difficult to compute. But if $\beta_f = \infty$, then Ψ is given by the maximum of the $\Psi_{m,n}$. Under this condition minimum bit-rate for a given Ψ is achieved when all $\Psi_{m,n} = \Psi$.

6.7 Optimization Method

Under the assumption $\beta_f = \infty$, the joint optimization of the quantization matrix reduces to the vastly simpler separate optimization of the individual elements of the matrix. Each entry of the perceptual error matrix $\Psi_{m,n}$ may be considered an independent monotonically increasing function of the corresponding entry $Q_{m,n}$ of the quantization matrix.

6.8 Optimizing QM for a given bit-rate

To obtain a quantization matrix that yields a given bit rate with minimum perceptual error Ψ we note that the bit rate is a decreasing function of Ψ , and it is a simple matter to estimate the requisite perceptual error.

7. APPLICATION TO SPACE IMAGERY

Image compression will play a vital role in the distribution of preview images of science data to scientists at distributed sites, especially in programs such as EOS and the Mission to Planet Earth (ref. 18). Due to the generally high performance and wide availability of the JPEG still image compression standard, we expect it to play an important role in this area. Since the JPEG standard includes the quantization matrix as part of the file, DCTune technology provides a method of optimizing the bit-rate/quality trade-off for each science image.

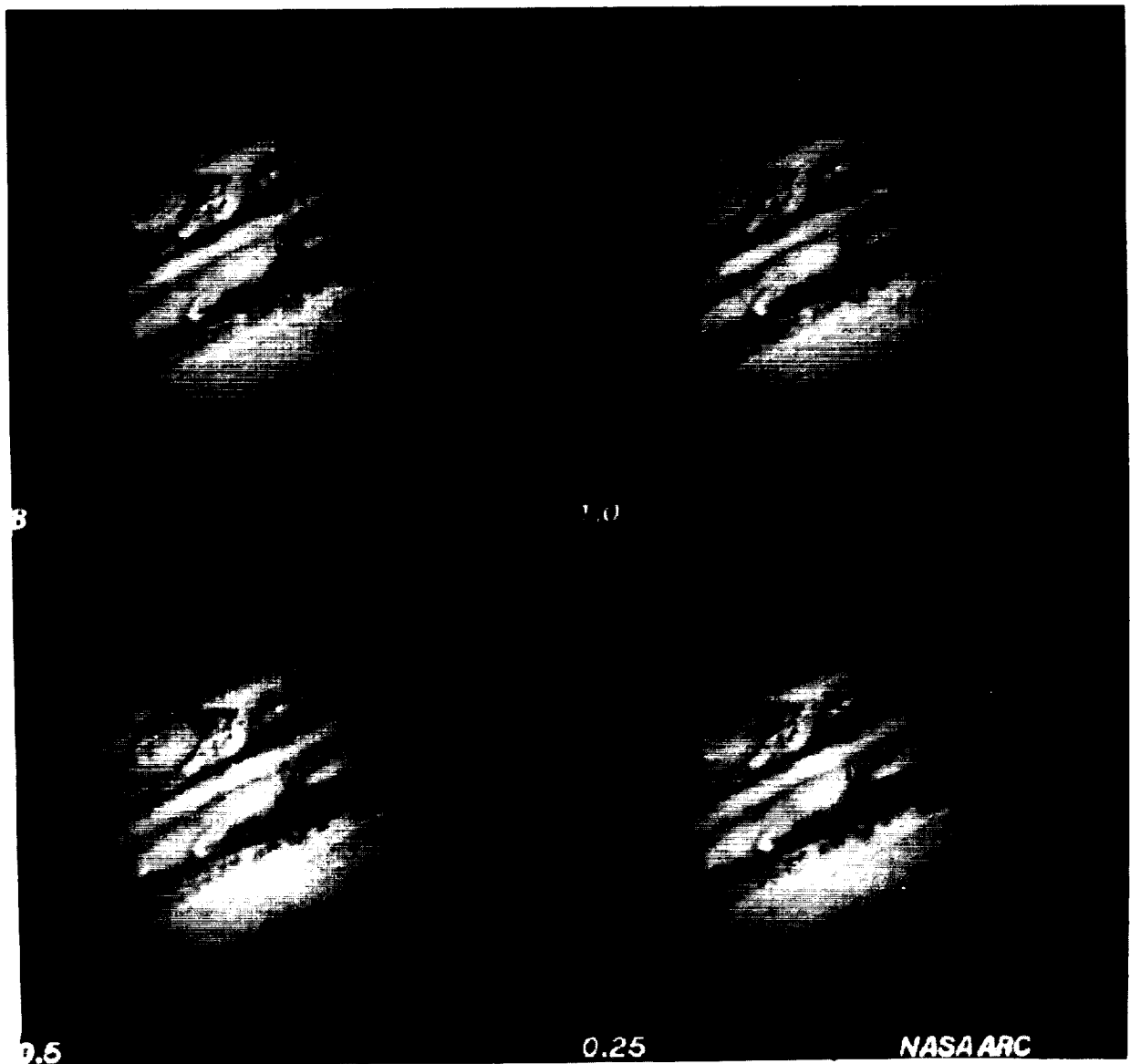


Figure 1. Voyager image of Jupiter compressed to 1.0, 0.5, and 0.25 bits/pixel, using optimal DCTune quantization matrices.

Lossy image compression based on the DCT may also play a role in the recovery of scientific imagery from spacecraft. The Galileo orbiter spacecraft is now on its way to Jupiter. Due to a malfunction of the main antenna, image data will be sent to earth over an auxiliary antenna with approximately 15,000 times lower bandwidth. Image compression will be used to partially compensate for the loss of bandwidth(ref. 19). In support of this effort, we have designed quantization matrices for use in the Galileo mission, based on application of DCTune technology to existing Voyager and Galileo images (ref. 20, 21). An example of DCTune algorithm applied to an image of Jupiter obtained by the previous Voyager mission is shown in Fig. 1. It shows the original and three levels of optimized compression: 1.0, 0.5, and 0.25 bits/pixel. In this example, the parameter values used were $a_T = 0.65$, $\beta = 4$, $w_{m,n} = 0.7$, display mean luminance $\bar{Y} = 65 \text{ cd m}^{-2}$, image greylevels = 256, $\bar{I}_{00} = 1024$. The viewing distance was assumed to yield 32 pixels/degree.

8. SUMMARY

We have shown how to compute a visually optimal quantization matrix for a given image. These image-dependent quantization matrices produce better results than image independent matrices. The DCTune algorithm can be easily incorporated into JPEG-compliant applications.

In a practical sense, the DCTune method proposed here solves two problems. The first is to provide maximum visual quality for a given bit rate. The second problem it solves is to provide the user with a sensible and meaningful quality scale for JPEG (or other DCT-based) compression. Without such a scale, each image must be repeatedly compressed, reconstructed, and evaluated by eye to find the desired level of visual quality.

9. ACKNOWLEDGMENTS

We appreciate the help of Heidi A. Peterson and Jeffrey B. Mulligan. This work was supported in part by the IBM Independent Research and Development Program and by NASA RTOP Nos. 506-59-65 and 505-64-53.

10. REFERENCES

1. Wallace, G. The JPEG still picture compression standard. *Communications of the ACM*. 34(4): 30-44, 1991.
2. Pennebaker, W. B. and J. L. Mitchell. "JPEG Still image data compression standard." 1993 Van Nostrand Reinhold. New York.
3. LeGall, D. MPEG: A video compression standard for multimedia applications. *Communications of the ACM*. 34(4): 46-58, 1991.
4. Peterson, H. A., H. Peng, J. H. Morgan and W. B. Pennebaker. Quantization of color image components in the DCT domain. *Human Vision, Visual Processing, and Digital Display*. Proc. SPIE. 1453: 210-222, 1991.
5. Peterson, H. A. "DCT basis function visibility in RGB space." *Society for Information Display Digest of Technical Papers*. Morreale ed. 1992 Society for Information Display. Playa del Rey, CA.
6. Peterson, H. A., A. J. Ahumada Jr. and A. B. Watson. The Visibility of DCT Quantization Noise. *SID Digest of Technical Papers*. XXIV: 942-945, 1993.
7. Peterson, H., A. Ahumada and A. Watson. "An Improved Detection Model for DCT Coefficient Quantization." *Human Vision, Visual Processing, and Digital Display IV*. Allebach ed. 1993 SPIE. Bellingham, WA.
8. Ahumada, A. J., Jr. and H. A. Peterson. "Luminance-Model-Based DCT Quantization for Color Image Compression." *Human Vision, Visual Processing, and Digital Display III*. Rogowitz ed. 1992 Proceedings of the SPIE.
9. Boynton, R. M. "Human Color Vision." 1979 Holt, Rinehart and Winston. New York.
10. Mullen, K. T. The contrast sensitivity of human color vision to red/green and blue/yellow chromatic gratings. *Journal of Physiology, Lond*. 359(381-400): 1985.
11. Watson, A. B. "DCT quantization matrices visually optimized for individual images." *Human Vision, Visual Processing, and Digital Display IV*. Rogowitz ed. 1993 SPIE. Bellingham, WA.
12. Watson, A. B. DCTune: A technique for visual optimization of DCT quantization matrices for individual images. *Society for Information Display Digest of Technical Papers*. XXIV: 946-949, 1993.
13. van Nes, F. L. and M. A. Bouman. Spatial modulation transfer in the human eye. *Journal of the Optical Society of America*. 57: 401-406, 1967.

14. Barlow, H. B. "Dark and light adaptation: Psychophysics." Handbook of Sensory Physiology. Hurvich and Jameson ed. 1972 Springer-Verlag. New York.
15. Legge, G. E. and J. M. Foley. Contrast masking in human vision. Journal of the Optical Society of America. 70(12): 1458-1471, 1980.
16. Legge, G. E. A power law for contrast discrimination. Vision Research. 21: 457-467, 1981.
17. Robson, J. G. and N. Graham. Probability summation and regional variation in contrast sensitivity across the visual field. Vision Research. 21: 409-418, 1981.
18. Jaworski, A. Earth Observing System (EOS) Data and Information System (DIS) software interface standards. AIAA/NASA Second International Symposium on Space Information Systems. AIAA-90-5075: 1990.
19. Cheung, K.-M. and K. Tong. Proposed data compression schemes for the Galileo S-band contingency mission. 1993 Space and Earth Science Data Compression Workshop. 3191: 99-109, 1993.
20. Watson, A. B. and A. J. Ahumada Jr. Preservation of photometric accuracy in ICT-compressed imagery. 1993.
21. Watson, A. B., A. J. Ahumada Jr. and M. J. Young. ICT quantization matrix design for the Galileo S-Band Mission. 1993.

11. APPENDIX

luminance quantization matrix	16	11	10	16	24	40	51	61
	12	12	14	19	26	58	60	55
	14	13	16	24	40	57	69	56
	14	17	22	29	51	87	80	62
	18	22	37	56	68	109	103	77
	24	35	55	64	81	104	113	92
	49	64	78	87	103	121	120	101
	72	92	95	98	112	100	103	99
chrominance quantization matrix	17	18	24	47	99	99	99	99
	18	21	26	66	99	99	99	99
	24	26	56	99	99	99	99	99
	47	66	99	99	99	99	99	99
	99	99	99	99	99	99	99	99
	99	99	99	99	99	99	99	99
	99	99	99	99	99	99	99	99
	99	99	99	99	99	99	99	99

Table 1. The default quantization matrices. The Q_{00} value is located in the upper left corner of each matrix.

Y' quantization matrix	15	11	11	12	15	19	25	32
	11	13	10	10	12	15	19	24
	11	10	14	14	16	18	22	27
	12	10	14	18	21	24	28	33
	15	12	16	21	26	31	36	42
	19	15	18	24	31	38	45	53
	25	19	22	28	36	45	55	65
	32	24	27	33	42	53	65	77
Cr quantization matrix	21	21	41	45	55	71	92	120
	21	37	39	38	44	55	70	89
	41	39	51	54	59	69	83	103
	45	38	54	69	80	91	106	126
	55	44	59	80	100	117	136	158
	71	55	69	91	117	144	170	198
	92	70	83	106	136	170	206	243
	120	89	103	126	158	198	243	290
Cb quantization matrix	45	43	103	114	141	181	236	306
	43	78	99	97	113	140	178	228
	103	99	130	138	150	175	212	262
	114	97	138	176	203	232	270	321
	141	113	150	203	254	299	347	403
	181	140	175	232	299	367	434	505
	236	178	212	270	347	434	525	619
	306	228	262	321	403	505	619	739

Table 2. YCrCb quantization matrices. The values in these matrices are obtained following the procedure described in Section 3. The $Q_{0,0}$ value is located in the upper left corner of each quantization matrix. As specified in the JPEG standard, the values have been rounded to the nearest integer. JPEG also requires that values in the quantization matrix be ≤ 255 .

VOICE AND VIDEO TRANSMISSION USING XTP AND FDDI

John Drummond

**Naval Command, Control and Ocean Surveillance Center RDT&E Division (NRaD)
San Diego, CA 92152-5000**

Edwin Cheng

**Naval Command, Control and Ocean Surveillance Center RDT&E Division (NRaD)
San Diego, CA 92152-5000**

Will Gex

**Naval Command, Control and Ocean Surveillance Center RDT&E Division (NRaD)
San Diego, CA 92152-5000**

ABSTRACT

The use of XTP and FDDI provides a high speed and high performance network solution to multimedia transmission that requires high bandwidth. FDDI is an ANSI and ISO standard for a MAC and Physical layer protocol that provides a signaling rate of 100 Mbits/sec and fault tolerance. XTP is a Transport and Network layer protocol designed for high performance and efficiency and is the heart of the SAFENET Lightweight Suite for systems that require high performance or realtime communications. Our testbed consists of several commercially available Intel based i486 PCs containing off-the-shelf FDDI cards, audio analog-digital converter cards, video interface cards, and XTP software. Unicast, multicast, and duplex audio transmission experiments have been performed using XTP and FDDI. We are working on unicast and multicast video transmission. Several potential commercial applications are described.

INTRODUCTION

Multimedia (voice, video, data, text, and graphics) distribution over high speed networks has many commercial applications which will revolutionize the way we use computers and networks. Several big corporations have already formed strategic alliances to explore new opportunities in this area.

We have been researching and experimenting for several years with high speed networks which utilize Fiber Distributed Data Interface (FDDI), and a high performance network protocol called Xpress Transfer Protocol (XTP). As multimedia increased in popularity, in both military and commercial world, we started to look at the possibility of using XTP and FDDI in voice and video transmission. We have performed many voice transmission experiments using XTP with several PC i486 machines connected via FDDI network. The results indicate voice transmission using XTP and FDDI have many advantages over traditional methods of voice transmission such as fault tolerance, high bandwidth, and data integration. Right now, we are performing video transmission experiments using XTP and FDDI.

XTP

The Xpress Transfer Protocol (XTP)[1] has been developed over the past seven years from a consortium of private industry, academia, and government to address many high performance and realtime issues that were lacking in previously developed transport and network protocols. Certain concepts from existing protocols (e.g. VMTP, GAM-T-103, Delta-t, NETBLT) were modified and combined with new ideas to form the basis for XTP. Experience and other ideas have added to its development to produce the current specification[2].

It is a protocol that spans the Network and Transport layers (layers 3 and 4) of the OSI 7 layer model and therefore has some interesting features due to the coupling of an end-to-end protocol with an intermediate, network protocol (e.g. bandwidth reservation of an intermediate resource by an end host or packet priority assigned by an end host and used by an intermediate router). Because the protocol does not specify policy but

supplies many mechanisms, it is very flexible and allows an application to determine the needed policy (e.g. reliability or best effort, connection or datagram, multicast or unicast). This flexibility allows applications to communicate efficiently without undue overhead from unneeded mechanisms or policies.

Some of the features of XTP that address efficiency are:

- Multicasting with some end-to-end reliability
- Rate control
- Selective retransmission
- Application control of acknowledgments
- Implicit connection setup with data
- Flexible error control (reliable or best effort)
- Flexible flow control (variable window or no flow control)
- Optional data checksum in a trailer (vice header)
- Connection ID exchange
- Header and trailer field alignment on 4 byte boundaries

XTP supports realtime systems in three areas by offering:

- Flexible communication paradigms
- Flexible degrees of reliability
- Message discrimination for scheduling

In the voice and video applications developed at NRaD, some of these mechanisms were used to increase the efficiency of the communications. The largest efficiency gain came from the use of multicast which allowed the source host to transmit a single voice or video packet over the network to a group of receivers. As this group of receivers grew, little additional overhead was required of the transmitter to distribute the voice or video to the larger group.

FDDI

The Fiber Distributed Data Interface (FDDI) is an American National Standards Institute (ANSI) standard based on token ring technology. Basically, it is a network of nodes connected by two fiber cables with a logical token circulating among the nodes and a signaling rate of 100 Mbits/sec. The FDDI architecture is fault tolerant and the network will remain operational if a single fault, such as a cable break, occurs. The FDDI features that are useful in the transmission of voice and video are:

- High bandwidth (100 Mbit/sec)
- Very low error rates (10^{-9} BER)
- Predictable token access (low jitter)
- Large packet size (4500 bytes)

Other features that are useful for Military purposes are:

- Fault tolerance
- No electromagnetic emissions
- No electromagnetic interference
- Notion of priority

EXPERIMENTS

Our testbed consists of several commercially available Intel based i486 PCs containing off-the-shelf components. Each network node PC is populated with Dual Attached Station (DAS) FDDI cards developed by Network Peripherals Inc. (model NP-EISA/2), 10-bit resolution audio analog-digital converter cards using Adaptive Differential Pulse Code Modulation (ADPCM) audio compression with 16 kHz sample rate developed by Antex Electronics (model VP 625), and 2-card set of video interface boards consisting of Visionary VIS-1A Joint Photographic Experts Group (JPEG) video compression hardware and Visionary VIS-2A frame grabber engine which perform the digital video capturing and compression functions (developed by Rapid Technology Corp). All PCs are connected by a dual ring fiber optic cable. The software consists of DOS 5.0 (developed by Microsoft), drivers for each of the hardware items mentioned above (developed by their respective companies) and XTP network software (developed by the Computer Networks Laboratory of the University of Virginia.) The primary and secondary storage of each node is 16 MB SIMM and 340 MB IDE type respectively.

The experiments we have performed follow two basic sequences of operation. The primary operation occurs on the transmitting side of the network communication connection. First task is to obtain video or audio analog data and to perform analog-to-digital conversion. Once completed this digitized information is then compressed using the appropriate data compression algorithm with the respective VLSI chipset (ADPCM for audio and JPEG for video) residing on the Commercial Off The Shelf (COTS) hardware. The compressed data is then formed into XTP network packets, by the application and XTP software, for transmission over the Fiber Optic network via the FDDI hardware in the node. The second set of operations is performed at the receiving node, beginning with the receipt by the Fiber Optic network and FDDI hardware in the node. These FDDI frames are processed by XTP and the resulting compressed data is delivered to the appropriate COTS hardware for decompression and the resulting data is output to either the screen in the case of video data or the speaker in the case of audio data. This completes the communication cycle.

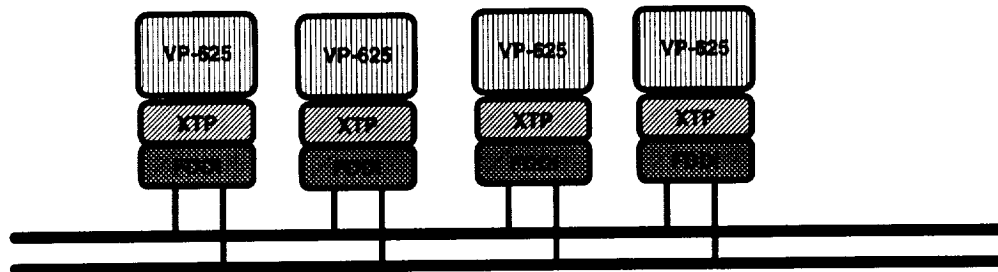


Figure 1. Schematic diagram of the Audio transmission experiment using XTP and FDDI

Voice Transmission

During the course of our audio experiments latency measurements were taken periodically. This measured latency was indicative of the time required for end to end (i.e. microphone-speaker) transmission. The results indicated a latency of approximately 25ms, which was well within the tolerable limits of perception by the human ear.

One-Way Delay	Effect of delay
>600ms	Incoherent
600ms	Barely Coherent
250ms	Annoying
100ms	Imperceptible (without network/original sample comparison)
50ms	Imperceptible (with network/original sample comparison)

Table 1: Effects of latency on human ear perception[3]

The latency tests were based upon XTP unicast mode communication link sending XTP network packets sized at 50 bytes each, and VP-625 A/D converter buffered by an array of bytes 1024 long. This schema provided a good basis for testing and analysis.

The ADPCM audio compression algorithm, which the Antex model VP-625 utilizes, compresses the sampled audio waveform to 4 bits thereby reducing the data size by over 50% compared to 10 bit PCM digitization. This compression allows for very low network bandwidth utilization. When operating in unicast mode, the average consumption of network bandwidth given a typical compressed audio packet is .5035 Mbits/sec. XTP unicast mode is a network communication utilizing 2 nodes, where one node acts as a transmitter and another node acts as a receiver. This bandwidth is increased to approximately 1.102 Mbits/sec when utilizing duplex mode communication. XTP duplex mode involves two nodes and each node acts as transmitter and receiver simultaneously. The experiments have included various aspects of XTP functionality such as option bits testing.

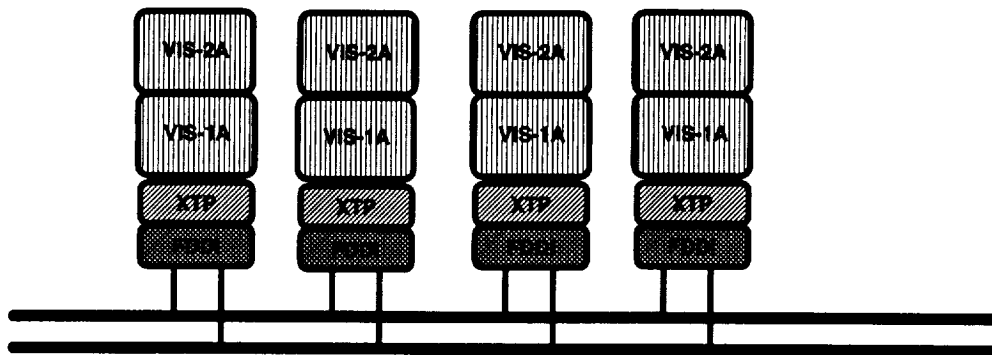


Figure 2. Schematic diagram of the Video transmission experiment using XTP and FDDI

Video Transmission

The realtime video we used originated from various sources such as: Cable News Network (CNN) broadcast acquired from satellite downlink; Video Camera; and Video Tape. These sources all followed the NTSC format and all were fed into the VIS-2A video frame grabber. Again, during our experiments, latency measurements were recorded periodically. This time the measured latency was representative of the time required for picture to picture (i.e. screen display to screen display) transmission. The outcome of these tests revealed a latency of approximately 50-60ms (less than 2 frames) given that our experiments were based upon utilizing NTSC standard input which is 30 fps. This small latency is very difficult to perceive, even with the source and destination display screens side by side. Table 2 represents some results of a study [3] on frame rates and their effects on human eyes. As can be seen, a jerky motion is perceived when successive frames are 67-83 ms apart. This is in excess of the latency in our experiment between a frame appearing on the source screen and the same frame appearing on the destination screen.

Frames per second	Effect on human eye
<10 fps	Frames appear disjoint
12-15 fps	Motion is jerky.
30 fps	Television quality
60-75 fps	High-motion discernible (HDTV)
90 fps	Limit of human eye perception

Table 2: Effects of frame rate on human eye perception[3]

These latency tests were also based upon XTP unicast mode communication link with XTP network packets sized at 3305 bytes each, and a video buffer of 16000 bytes. The JPEG video compression chipset utilizing

the Huffman encoding scheme provided us with 2 to 4 times data reduction thus greatly reducing the network bandwidth requirements for our realtime video communication experiments. A typical XTP unicast communication session utilized approximately 3 Mbits/sec to 6 Mbits/sec bandwidth. The XTP multicast sessions were recorded to also within the 3 Mbits/sec to 6 Mbits/sec range of network bandwidth consumption. The results of the XTP multicast bandwidth consumption is revealing in that it is within the range of the typical XTP unicast network utilization despite the fact that unicast is a 1 to 1 session and multicast is 1 to N network communication

POTENTIAL COMMERCIAL APPLICATIONS

Medical Image Transmission

Digitized medical image transmission over high speed network and using high performance network protocol will save a lot of time for doctors and laboratories to diagnose symptoms of patients. For example, after a X-ray laboratory takes the image of a patient's skull, the image can then be transmitted via the FDDI to a doctor's office and displayed in his monitor. Also it can be sent to several locations at the same time. This will save time for film processing, filing, mailing and paperwork. Both XTP and FDDI are apt for this task since medical image transmission requires high bandwidth and no error rate. Other information such as billing, documents, and patient's data can also be transmitted using the same network. Hence the amount of medical paperwork will be reduced.

On the basis of the video experiments we performed, the potential of applying our technology to the medical image transmission is good and is worth more researches.

Teleconferencing

Business organizations have been using teleconferencing for many years already. There are several advantages of using teleconferencing such as reducing travel expenses, travel time, and schedule problems. However teleconferencing has been limited by the speed of technology development. Low speed networks, low performance network protocols, and low performance hardware have created latency in voice transmission and limited the number of video frames to be transmitted.

The incorporation of XTP and FDDI into a teleconferencing network will greatly improve the performance of the teleconferencing facilities. Voice latency will be reduced and the number of video frames can be increased.

In our voice experiments, we have finished duplex, unicast, and multicast. In our video experiments, we are working on unicast and multicast. These experiments have demonstrated the feasibility of teleconferencing using XTP and FDDI. Later on, we will incorporate the audio and video experiments together in a coherent manner.

Right now our voice and video experiments are only in DOS environment using PC 486 machines. We will incorporate XTP and FDDI into RTMach environment using PC i486 machines in the near future. RTMach is a high performance realtime operating system developed by Carnegie Mellon University. After incorporation, more experiments that are not feasible in DOS such as priority experiments can be conducted in RTMach environment.

Voice and Electronic Mail in High Traffic Networks

Voice and Electronic mail have been used in business, research organizations, and government for many years. These kinds of communication provide a convenient way for users between different building and even different countries. As the number of users and the number of messages increase several times each year, the amount of network traffic also increases at a tremendous speed. This creates a need for higher performance

network and network protocols to satisfy the demanding network traffic. The use of XTP and FDDI may provide a solution. However, more research need to be done in order to incorporate existing networks to XTP and FDDI.

Remote Data Acquisition

Data Acquisition has many applications in industry. Some of these applications such as air tunnel analysis require large amount of data acquisition and transmission through computer network. This will inevitably create a bottleneck effect on the network. The use of XTP and FDDI provides a potential solution to improve the performance of the network traffic and thus facilitate the data flowing.

CONCLUSION

After performing the audio and video experiments in our testbed, the results have indicated the use of XTP and FDDI on multimedia transmission is feasible. Several potential commercial applications are described. As multimedia will become an important industry in our nation, more research in multimedia transmission is needed in order to provide an edge over our competitors.

We will cooperate with Carnegie Mellon University and University of Virginia to conduct more experiments in the future to utilize more powerful computers, operating systems, audio, and video equipment. These experiments will provide higher performance multimedia transmission, benchmarks, research data, and new technology to industries.

REFERENCES

- [1] W. Timothy Strayer, Bert Dempsey, Alfred Weaver, "XTP: The Xpress Transfer Protocol," Addison-Wesley Publishing Co, Inc., 1992.
- [2] "XTP Protocol Definition," Protocol Engines Inc. Report PEI 92-10, 1992.
- [3] Amit Shah, Don Staddon, Izhak Rubin, Aleksandar Ratkovic, "Multimedia Over FDDI," IEEE Proceedings, Sept, 1992, pp 13-15

**AN INTELLIGENT INTERACTIVE VISUAL DATABASE MANAGEMENT SYSTEM FOR
SPACE SHUTTLE CLOSEOUT IMAGE MANAGEMENT**

Dr. James M. Ragusa,* Dr. Gary Orwig, Michael Gilliam,*
David Blacklock,* and Ali Shaykhian*
University of Central Florida
Orlando, FL 32816-0112**

*** College of Engineering
Department of Industrial Engineering
and Management Systems**

**** College of Education
Department of Educational Services**

ABSTRACT

This paper provides status on an applications investigation of the potential for using an expert system shell for classification and retrieval of high-resolution digital, color space shuttle closeout photography. This NASA funded activity has focused on the use of integrated information technologies to intelligently classify and retrieve still imagery from a large, electronically stored collection. In this paper a space shuttle processing problem is identified, a working prototype system is described, and commercial applications are identified. A conclusion reached is that the developed system has distinct advantages over the present manual system and cost efficiencies will result as the system is implemented. Further, commercial potential exists for this integrated technology.

INTRODUCTION

Research under way at the University of Central Florida (UCF) College of Engineering's Intelligent Multimedia Applications Laboratory (IMAL) is directed at the investigation of the feasibility and integration of knowledge-based (expert) systems technology to facilitate the use of a large microcomputer-based collection of high-resolution digital photographic images. Specifically, a NASA-sponsored research project focuses on the collection, compression, classification, storage, retrieval, and transmission of high-resolution Kennedy Space Center (KSC) space shuttle ground processing still color photographs ([1],[2],[3]). The following briefly describes this application.

Application Background

For each space shuttle ground processing cycle, approximately 35,000 chemically developed 8 x 10 inch still photographs (including copies) document all significant pre-launch and post-landing activities. These pictures are an integral part of NASA's quality control and reliability program that assists space shuttle systems engineers in verifying "go for launch" and the condition of the orbiter after landing.

Photographs of key subsystem elements of space shuttle systems (orbiter, main engines, external tank, solid rocket boosters, and ground support equipment) are contained in a permanent collection of almost two million pictures. Included in these detailed photographs are pumps, rocket nozzles, connectors, cabling, control panels, mechanical assemblies, protective tiles, and valves. Presently, this collection is manually catalogued and stored in notebook binders, file cabinets, and storage boxes.

When a KSC systems engineer, or other user, needs an individual photograph or a group of related photographs, a search is performed by expert staff specialists. The present system allows

manual searches by mission number, key space shuttle elements, work authorization numbers, or by unique image accession number. Retrieval is difficult or impossible without this information. Frequently, the needed photo or set is found only after a time-consuming search. This effort is repeated at other NASA and contractor locations around the country.

Retrieved processing photographs have been used to answer critical configuration processing questions on numerous occasions. They have also eliminated the need for expensive space shuttle disassembly or inspection in closed or hazardous locations. As should be evident, the cost of misplaced or lost shuttle photographs is very high when compared to a study indicating that even a misfiled document in an average business costs \$125, and each lost document costs \$350-\$700 [4].

Because of the importance of these photographs, NASA-KSC has sponsored research to investigate more cost-effective methods of system improvement, including the use of expert systems for the critical classification and retrieval phases of the Image Database Management System (IDBMS) life-cycle process.

Knowledge-Assisted System Enhancement

Theoretically and practically, the use of knowledge-based, or expert, systems has potential to positively impact NASA image classification and retrieval. This is important because of the large quantity of photographs involved and their difficult to describe technical orientation. In the NASA-KSC environment, staff specialists have, over the years, developed heuristics for image classification and retrieval tasks not commonly known or documented. For example, experienced classification specialists have learned to recognize which objects in technically oriented space shuttle subsystem pictures are important and which are not. Retrieval specialists also develop an understanding of the logical sequence systems engineers and quality technicians use to location required images. This knowledge is not possessed by less experienced, substitute, or newly hired classifiers or retrievers who replace those that are sick, transferred, or retired.

Another researcher shares the view that an expert system, functioning as a decision-aid to help inexperienced users, can improve the effectiveness and efficiency of image classification and retrieval for this class of users [5]. Importantly, several other researchers ([6],[7],[8]) agree that such on-the-job task domain expertise provides an ideal environment for expert systems application success.

PROTOTYPE DEVELOPMENT

A prototype that integrates expert system technology with an IDBMS has been developed for multi-mission NASA space shuttle color image classification and retrieval tasks. The system consists of two modules, one for image classification; the other, for image retrieval. They are named NAPSAC for NASA PhotoGraph System to Aid Classification; and, PRAISE, for Photo Retrieval And Identification System Expert [3].

Knowledge acquisition was accomplished by UCF-IMAL researchers who met with KSC classification and retrieval domain experts. Numerous interviews were conducted using widely-recognized knowledge acquisition and structuring methods ([8],[9],[10]). Resultant image processing paradigms, decision criteria, descriptor attributes, and heuristics were structured, modeled, and converted into code using LEVEL5 OBJECT (an object-oriented expert system development tool). This expert system shell was chosen because of its capabilities for:

- Developing rule-based and object forms of knowledge representations.
- Manipulating and displaying user-friendly bitmapped images.

- Creating hypertext and object-oriented interfaces.
- Calling external programs needed to access relational databases (in this case, dBASE III PLUS) and image storage devices.
- Integrating with IBM microcomputer hardware and Microsoft Windows software.

Image Classification

The classification module uses the expert system development shell's hypertext and object-oriented features. For example, during image classification, a cataloger views a hierarchy of bitmapped object images of space shuttle elements on various screens as shown in Figure 1. The screen at the first level pictures the entire space shuttle configuration. "Hyperregions" (shown as dashed boxes) outline major shuttle elements (objects)--the orbiter, external tank, solid rocket boosters, and ground support equipment. Screen notes (not shown) provide user instructions. If the cataloger uses a mouse selection device to point-and-click on the orbiter hyperregion, the orbiter is displayed on a second level screen where the bitmapped orbiter object and its forward fuselage, payload bay, and tail section hyperregions are outlined.

After further point-and-click selection (for instance, to the forward fuselage hyperregion), a third level screen displays the forward fuselage with flight deck, middeck, and lower deck hyperregions. A fourth level selection (for example the flight deck of the forward fuselage) allows final object selection of specific subsystems (for example, hand controller or pilot control panel). Only four levels of object hyperregion image displays are implemented in this prototype classifier system. The reason being that four levels of classification appears more than adequate, after user testing, to accommodate image processing paradigms, decision criteria, descriptor attributes, heuristics, and data requirements which were identified during the knowledge acquisition process. No required information appears lost with this level of specificity.

A database record is automatically created as a result of the cataloger's manual input of selected image data and navigation through the bitmapped object images. Image record information includes work authorization numbers, vehicle and mission information, photographic image data, and a sequential image number which is assigned by the expert system. A photo description area is available for special cataloger notes or comments. Aspect and sub-aspect entries are entered automatically as a result of point-and-click selection. Image record database information is stored on a hard disk, and the images are stored separately in a compressed mode on a mass storage device.

Image Retrieval

The retrieval module allows engineers and other casual users to locate and display space shuttle processing images that have been previously classified and stored. The image retriever expert system works much as the classifier does except that it does not create database records.

During the image retrieval process, the user is first allowed to directly input selected attribute values such as a KSC photo or work authorization number, if known, for immediate image retrieval. If this information is unknown, the vehicle name, mission number, and/or photo acquisition information are identified. Next, the screen sequence of bitmapped space shuttle images described for the classifier and illustrated in Figure 1 is used (with point-and-click operation). Attribute values (for example, orbiter, forward fuselage, flight deck) are automatically constructed by the expert system shell to serve as search query criteria for image retrieval. Like the classifier, four screen displays levels in the prototype system were found by KSC users to provide needed retrieval capabilities.

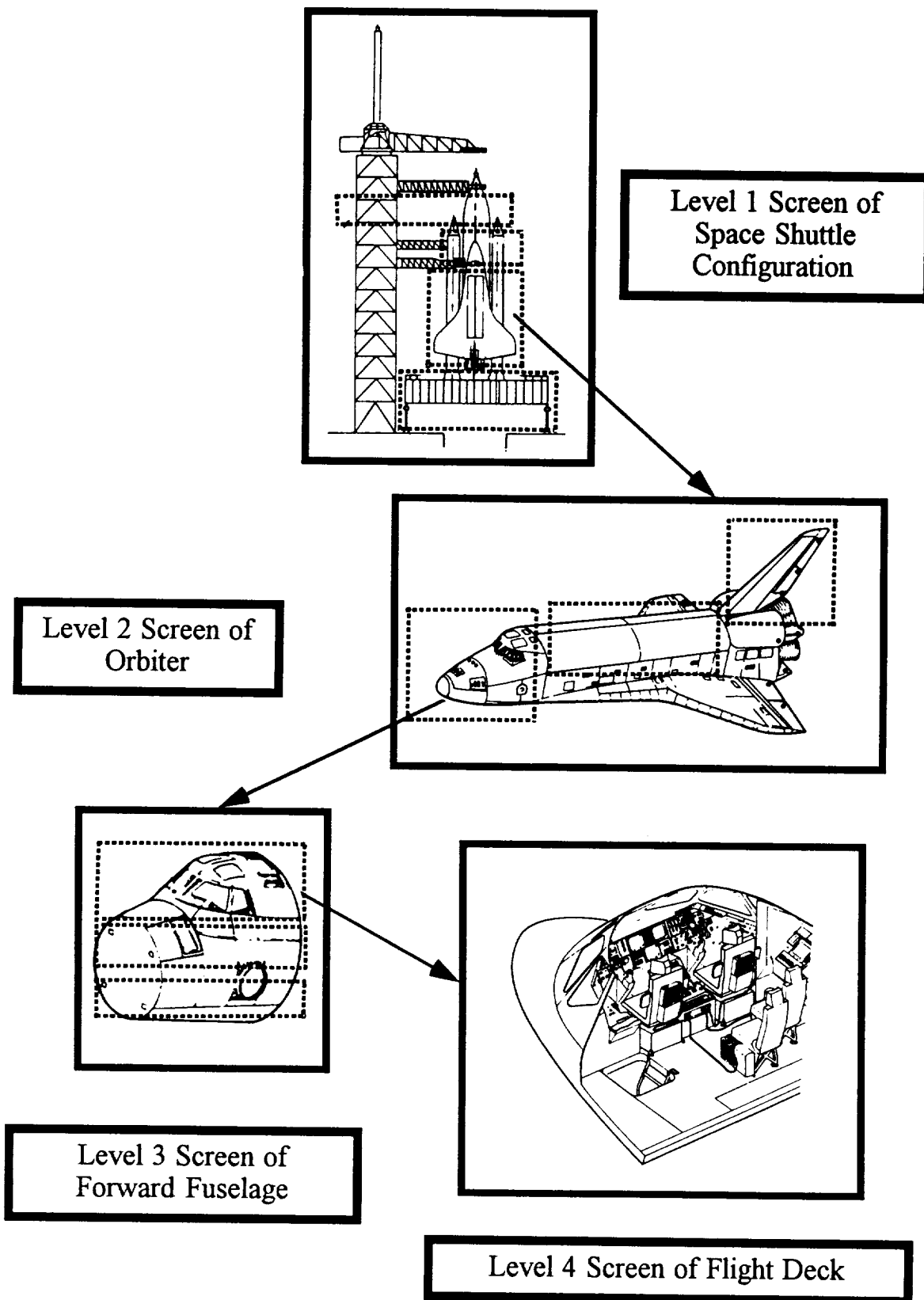


Figure 1. Diagram of classifier/retriever screens with hyperregions indicated.

After the query has been structured, the database record is searched using the known attribute values. Any "hits" are indicated in a counter screen display for the system user. These records are then used to retrieve thumbnail images from the image mass storage device for display. These surrogates allow the user to determine image relevancy.

Using the mouse, a user selects a desired thumbnail image. The system automatically uses the image number pointer in the database record to locate the desired image from data storage. Image object attribute values (e.g., KSC photo number, mission number, date, etc.) from the database record are first displayed. Then, the full screen, full color image is shown after decompression. This technique follows the cognitive model used by domain experts as well as that described by other researchers [11]. After viewing the image and its associated data, a user can choose to return to the screen of thumbnail images for additional selections or may choose to exit the session. Help text screens and instructional notes are available throughout the system.

The retriever and classifier prototype systems enable faster retrieval and classification of still photographs (when compared to the old system). Both eliminate or reduce many procedural steps, and the point-and-click feature greatly reduces many opportunities for human keyboard error. The expert system modules also enforce a standard terminology and single database repository, for the classification and retrieval of KSC image records. These improvements have the potential to eliminate or greatly reduce temporal classification variations, ensure consistency in classifier and uses orientation, and provide a more user-friendly environment.

PC-LAN VERSION

The prototype classification and retrieval modules work well for a one-user, single location system. However, space shuttle ground processing activities are spread over twenty-five square miles in at least six different KSC facilities. In addition, during pre-launch preparation and post-landing operations, engineers at locations throughout the country have responsibility for evaluating various space shuttle elements. A networked IDBMS is required to support the needs of these users.

A generalized PC-based local area network (LAN) prototype IDBMS design is shown in Figure 2. It is consistent with another researcher's assertion that there are advantages to an open architecture which enables seamless image and data access through the use of powerful servers and related desktop technology [5]. This architecture is organized into four primary modules: production station, file server, local user workstations, and remote workstations. The first three modules are connected by a high speed (16 megabits per second), token ring LAN network running under Novell Netware control. The remote workstations (the fourth module) will eventually be linked to the file server via high capacity communication links.

The purpose of the IDBMS production station is to input photographic images, update the record database using the classifier expert system, and compress images. Images can be captured by using either a high-resolution photographic color scanner, a television camera, or analog or digital still video electronic cameras.

The knowledge-assisted classifier is used to create a database record of the captured image during the acquisition process. Digital images are entered into the system directly from the photographic flatbed scanner or a digital still video system. An alternate method is to input analog images using a television camera (for 8 x 10 inch photographs) or an analog still video camera. For analog image digitization, a commercial analog-to-digital board in the production station microcomputer is used to create individual digital image files. Each of these files is then compressed

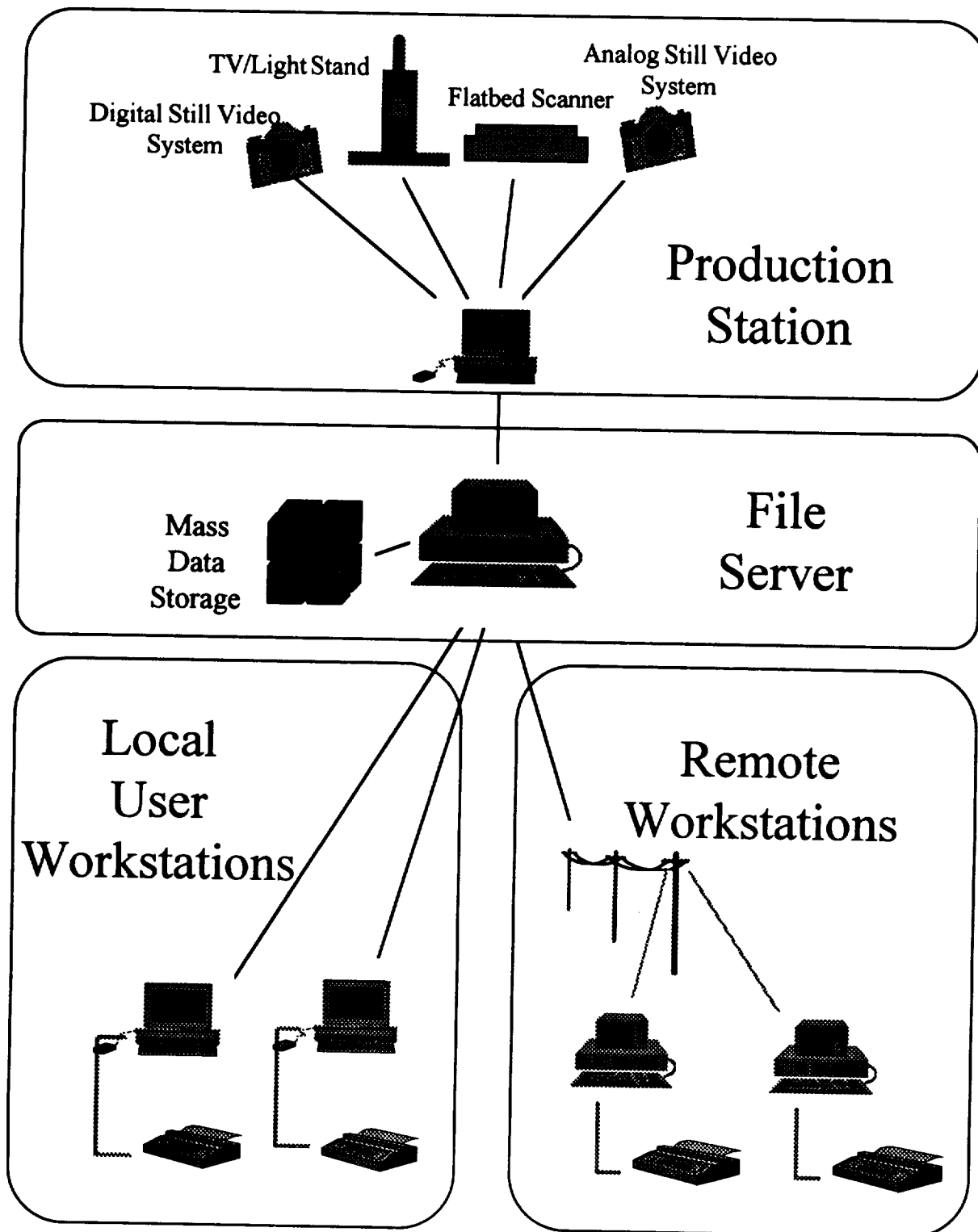


Figure 2. PC-LAN IDBMS.

by a JPEG-compatible board also located in the microcomputer. Compression ratios range from 30:1 to 35:1, depending on image complexity.

After image capture, analog-to-digital conversion (if required), and compression at the production station, image files are transmitted to the mass storage device attached to the file server. A one gigabyte erasable optical disk drive is used for mass image data storage for this prototype system. A higher capacity image storage device, in the terabyte range, would be required for an operational system. The database records used to locate images are stored as a database file on the file server's separate internal hard disk.

Images are retrieved for engineering and quality analysis at user workstations. The retrieval expert system, resident at each workstation, is used to build a database query for image retrieval and display. A query usually results in multiple hits which are displayed as bitmapped thumbnail pictures for user review and further selection. The desired image is then decompressed (using a JPEG-compatible board in each workstation) in real-time and displayed on a large high-resolution (1024 x 768 pixel), true color (24-bit) monitor. The image is printed, if needed, using a digital high-resolution color printer.

The final module in this architecture consists of remote workstations. They function the same as local KSC user stations for image retrieval and display, except that they are at various locations throughout the United States. During remote operation, engineers at these distant locations would access the KSC file server records database using the knowledge-assisted retrieval expert system resident on each workstation. In operation, compressed image files, retrieved from the image mass data storage unit, would be transmitted via existing high-speed transmission links from KSC to various user locations.

IMPORTANCE AND COMMERCIAL POTENTIAL

The paradigm shift of using PC-based networked digital image data and knowledge-based classification and retrieval instead of hard copy prints and manual image management is anticipated to result in a significant improvement in closeout image productivity and cost savings for NASA-KSC space shuttle processing operations. It is estimated that cost savings/avoidance of 50% are possible as a result of implementation of an interactive digital image management system using knowledge-based classification and retrieval.

Also significant is the commercial potential for use of these low-cost integrated technologies in a variety of private and public sector environments and applications. Potential environments include education and training, travel and tourism, commercial and residential real estate, medicine and dentistry, survey and construction, advertising and sales, and presentation and entertainment. Specific media requiring improved image management systems include education and training materials, medical X-ray and scanned data, satellite and aircraft imagery, high quality photographic collections, advertising and related graphic inventories, visual product images, and executive and corporate presentation materials. Important is the fact that developed software technologies are fully compatible with many existing private and public sector PC-based client-server computer environments. Implementation of visual database technologies will require application system analysis and design, domain-specific software development, and systems integration. As a result, potential exists to stimulate growth in a variety of small business computer service organizations and in the development of generic/industry specific software programs.

SUMMARY AND CONCLUSIONS

This paper has briefly reviewed features important to high-capacity photographic image data capture, classification, compression, storage, retrieval, and display. Also described was a NASA-KSC space shuttle ground processing prototype IDBMS under development which provides knowledge-based assistance for image classification and retrieval. Finally, a design for a networked PC-LAN IDBMS was presented. A conclusion reached from reviews of the prototype system is that it has distinct advantages over the present manual system and cost efficiencies will result as the system is implemented. Further, commercial potential exists for this integrated technology.

ACKNOWLEDGEMENT

This research project is sponsored under a NASA-KSC/UCF Cooperative Agreement (NAG-10-0058). Appreciation is expressed to NASA-KSC's Nancy Sliwa, Astrid Heard, and William Helms of the Advanced Projects Office for their vision and support.

REFERENCES

- [1] Ragusa, J.M. and Orwig, G, "Expert systems and imaging: NASA's start-up work in intelligent image management," Expert Systems, vol. 2, no. 3, 25-30, Winter 1990.
- [2] Ragusa, J.M. and Orwig, G., "Attacking the information access problem with expert systems," Expert Systems, vol. 2, no. 4, 26-32, Spring 1991.
- [3] Ragusa, J.M. and Wielgos, R., "Using expert systems to interface relational and object-oriented databases for a NASA space shuttle applications," Heuristics: The Journal of Knowledge Engineering, vol. 4, no. 3, 1-10, Fall 1991.
- [4] Coopers & Lybrand, Information & Image Management: The Industry & the Technology. New York: Author, 1987.
- [5] Thompson, D., "Imaging meets expert systems," AI Expert, vol. 6, no. 11, 24-32, Nov., 1991.
- [6] Harmon, P., Maus, R., and Morrissey, W., Expert Systems: Tools & Applications. New York: John Wiley, 1988.
- [7] Liebowitz, J., An Introduction to Expert Systems. Santa Cruz, CA: Mitchell Publishing, 1988.
- [8] Mockler, R.J. and Dologite, D.G., Knowledge-Based Systems: An Introduction to Expert Systems. New York: Macmillan, 1992.
- [9] McGraw, K.L. and Harbison-Briggs, K., Knowledge Acquisition: Principles and Guidelines. Englewood Cliffs: Prentice Hall, 1989.
- [10] Scott, A.C., Clayton, J.E., and Gibson, E.L., A Practical Guide to Knowledge Acquisition. Reading: Addison-Wesley, 1991.
- [11] Besser, H., "Visual access to visual images: The U Berkeley image database project," Library Trends, vol. 38, no. 4, 787-798, Winter 1990.

THE TRUSTWORTHY DIGITAL CAMERA: RESTORING CREDIBILITY TO THE PHOTOGRAPHIC IMAGE

Gary L. Friedman
 Technical Group Leader
 Advanced Information Systems
 Jet Propulsion Laboratory
 California Institute of Technology
 Pasadena, CA 91109

Introduction

The increasing sophistication of computers has made digital manipulation of photographic images (as well as other digitally-recorded artifacts, such as audio and video) incredibly easy to perform and, as time goes on, increasingly difficult to detect. Today, every picture appearing in newspapers and magazines has been digitally altered to some degree, with the severity varying from the trivial (cleaning up "noise" and removing distracting backgrounds) to the point of deception (articles of clothing removed, heads attached to other people's bodies, the complete rearrangement of city skylines). As the power, flexibility and ubiquity of image-altering computers continues to increase, the well-known adage that "the photograph doesn't lie" will continue to become an anachronism.

A solution to this problem comes from a concept called Digital Signatures, which incorporates modern cryptographic techniques to authenticate electronic mail messages. [1] [2] ("Authenticate" in this case means you can be sure that the message has not been altered, and that the sender's identity has not been forged.) The technique can serve not only to authenticate images, but also to help the photographer retain and enforce copyright protection when the concept of "electronic original" is no longer meaningful.

Background on Digital Signatures

The concept of a digital signature builds upon a recent encryption technique called "Public Key Encryption" [3]. Older encryption/decryption schemes require that both the sender and receiver possess the same secret "key": the sender uses the key to transform the text message into ciphertext, and the receiver uses the same key to perform an inverse transformation on the ciphertext, revealing the original text message. If the correct key transforms the ciphertext into unreadable garbage, it is reasonable to conclude that either the wrong key is being used, the message has been altered, or the sender has been impersonated by someone ignorant of the correct key. The historic drawback to

this secret key encryption scheme has been in the secure distribution of keys; key disclosure must occur out-of-band, either transmitted via an expensive alternate path or arranged when sender and receiver were proximate.

Public key encryption techniques differ in that they enable the recipient of a message to decrypt it using a key that is different from the one used by the sender to encrypt it. All public key cryptography is based on the principle that it is easy to multiply two large prime numbers together, but extremely difficult (taking perhaps centuries using today's supercomputers) to work backwards and uncover the factors that could have been used to generate the resulting number.

Public Key Encryption employs two different keys: a private key, which is held by the more security conscious party, and a corresponding public key, which need not be kept secret. The public key is generated based upon the private key, making the pair unique to each other.

The public key scheme is illustrated in Figure 1 and works as follows: to send a secret message that only the recipient can read, the recipient would first make his/her public key known to the sender through any non-secure medium, such as a letter, a telephone conversation, or a newspaper ad. Anyone wishing to send a secure message would encrypt the message using this public key and send it to the recipient. The recipient, having sole possession of the corresponding private key, is the only one able to decrypt the message. The need to transmit a secret key that both parties must possess beforehand has been eliminated. The tradeoff in this case is that, although only the recipient can read the message, anyone who obtains the public key can send a message with anonymity.¹

¹ The described scenario can also be used as the first step in a process of exchanging secret keys to allow for conventional secure message transmission, eliminating any of the drawbacks of the one-way authenticatability. [1], [4]

The process described above can also be implemented "backwards" to great advantage. In a second scenario, it is the sender who maintains possession of the private

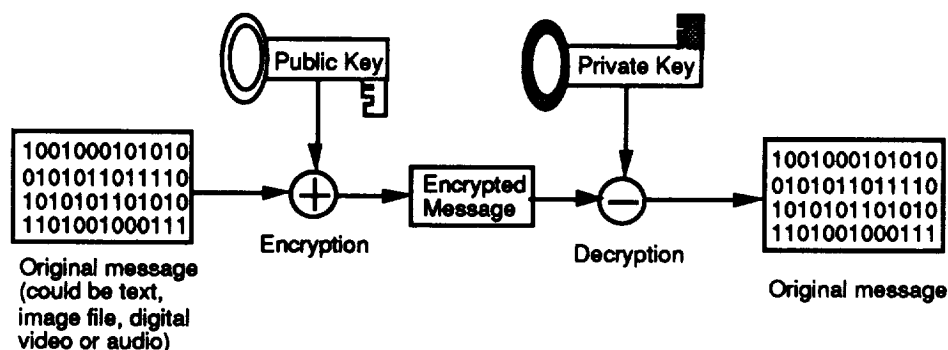


Figure 1: With Public Key Encryption, the encoding (public) key and decoding (private) keys are different, and it is computationally difficult to derive one given the other. To send a message that only the receiver can read, the non-secret public key is used to encrypt; the secret private key is used to decrypt. Encrypting with the private key forgoes confidentiality in favor of authenticity: if the public key can decode it, then only the one holding the private key could have generated the message.

key, and anyone who has the widely disseminated corresponding public key could decrypt this message. Although this procedure no longer performs the traditional function of encryption (which is to provide confidential communication between two parties), it does provide a way to insure that messages have not been forged: only the private key could have produced a message that is decipherable by the corresponding public key.

This gives us the foundation for message authentication: if the private key remains private, then *only* the private key holder can produce messages decipherable by the public key. Furthermore, it is extremely difficult to reverse-engineer the public key and ascertain the original private key. Without knowledge of the private key, a counterfeit message cannot be forged.

Digital signatures build upon these public key cryptographic techniques and allow you to authenticate the contents of the message as well as the identity of the sender, without obscuring the original message. The signatures are produced by creating a *hash*² of the

²A hash is a mathematical function which maps values from a large domain into a smaller range. For example, a checksum is a simple kind of hash. A more complex example of a hash algorithm would involve dividing a binary file into a collection of, say, 16 Kilobit pieces and performing a cumulative Exclusive OR function between successive pieces produces a simple 16 Kilobit "hash" which is smaller than the original file yet is practically unique to it. (Many more complex and secure transformations are also possible.) Changing a single bit in the original message produces a very different hash output; and

original plaintext message, and then encrypting the hash using the sender's private key, as shown in Figure 2. The result is a second digital file (referred to as a *signature*) which accompanies the original plaintext message. To emphasize: THE ORIGINAL MESSAGE IS UNTOUCHED; only the message's hash is encrypted. This way the original file can be read by all, yet if you wish to authenticate it you can decrypt the message's unique digital signature using the public key. If the decrypted digital signature and an independent hash on the file in question match, both the integrity of the message and the authenticity of the sender can be assured.

This digital signature technique is very general; it can be applied not only to 1-dimensional symbolic text (such as electronic mail) but also to any n-dimensional digital pattern (such as digital video, digital audio, and/or digital holograms).

Digital Cameras

Standard digital cameras are filmless; they sense light and color via an electronic device (such as a Charge Coupled Device (CCD)), and produce as output a computer file that describes the image using 1's and 0's arranged in a meaningful, pre-defined format. Often this digital image file is stored on a small mass-storage medium inside the camera itself (such as floppy disk or magneto-optical disk) for later transference to a large computer. Alternatively, the image file can be sent directly to the computer via a transmission medium. Once inside the computer it then can be read and then easily altered in any number of different ways.

In the proposed digital camera (Figure 3) we wish to authenticate the initial image file as it emerges from the camera. To accomplish this, the camera produces two output files for each captured image as shown in Figure

reverse engineering a message so it will have a given hash value and also make sense to the reader is virtually impossible. A digital signature can then be created by encrypting the hashing output using the sender's private key.

4: the first is an all-digital industry-standard file format representing the captured image. The second would be an encrypted "digital signature" produced by applying

originally produced. If on the other hand at least a single bit is different, the two hashes will not even closely match and the image's integrity will not be affirmed.

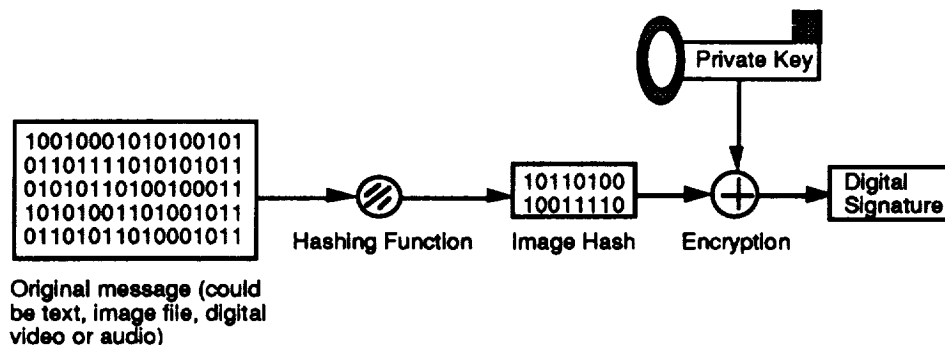


Figure 2: The Digital Signature is created by producing a complex checksum called a "hash", which is then encrypted using the private key. Attempting to forge this signature without knowledge of the private key would take decades using today's supercomputer technologies.

the camera's unique private key (embedded within the camera's secure microprocessor) to a hash of the captured image file, using the procedure described in [4]. It is the responsibility of the user to keep track of both files once they leave the camera, since both are required to authenticate the image.

Once the digital image file and the digital signature are generated by the camera and stored in computer memory, the image file's integrity can later be affirmed by using a public key decoding program, which can be freely distributed to users and certification authorities via conventional software distribution techniques. This verification program (illustrated in Figure 5), which has no knowledge of either the public or private keys, takes as input the digital image file in question, its accompanying digital signature file, and the public key which is unique to the originating camera. (It is perfectly reasonable to have the public key double as the camera's serial number.) The program then calculates its own hash on the digital image file (the hashing algorithm need not be kept a secret), and uses the public key to decode the digital signature to reveal the hash originally calculated by the camera at the time the image was taken (Figure 6). If these two hashes match, it is certain to any required degree that the digital image in question is indeed identical to what the camera

techniques (such as audio cassette tape or the NTSC encoding on today's video tape formats) or non-corrected digital formats (such as the popular audio compact disc (CD), which is so unreliable that CD player manufacturers now employ "over-sampling" to combat the problem of missed bits) introduce a large amount of errors upon playback that are normally imperceptible to the viewer or listener, but are intolerable for the purposes of image authentication.

Measures of Protection

The scheme as described above is resistant to forgery attempts since the secret private key (which is known

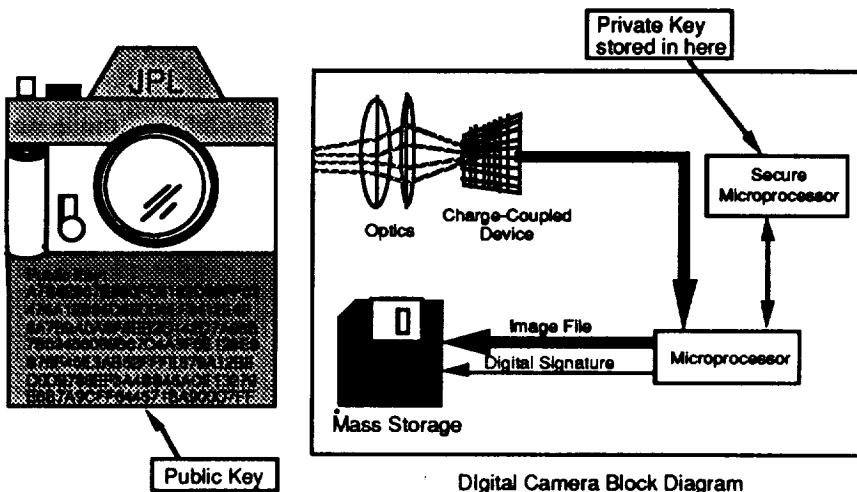


Figure 3: The Trustworthy Digital Camera starts with a digital sensor instead of film, and delivers the image directly in a computer-compatible format. The secure microprocessor responsible for the encryption of the digital signature is programmed with the private key at the factory. The public key necessary for later authentication appears in the image's border as well as on the camera body.

only by the camera's manufacturer) is embedded in a probe-proof microprocessor which itself is deeply integrated into the camera's system (Figure 1). Even if some adept pirate were to dissect the camera and replace

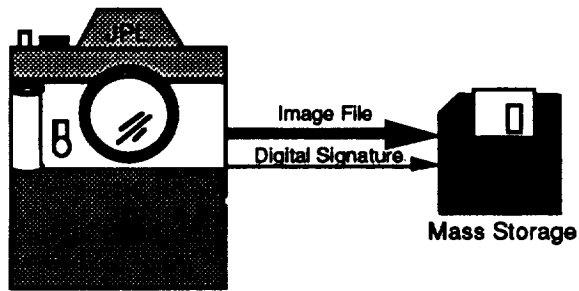


Figure 4: When a single photo is taken, two files are produced: a standard digital image file, and an encrypted digital signature. The files can be stored on a variety of media, such as a Write-Once-Read-Many (WORM) CD or the computer's mass storage device. The image can then be accessed and used just as any other computer data.

the chip with one containing a homebrew key, the digital signature produced thereafter would not be decodable by any public key published by the manufacturer.

The advantages to freely distributing the verification software and valid public keys are great; with the software freely available verification can become commonplace and routine. No special certification authority need be called in for routine checks, no fees are required, no big fuss is made, and no bad-faith climate amongst the parties involved need be created as a result of being challenged. But the mass distribution of verification software does carry one danger: it would be easy for someone to create a bogus program that looks, behaves, and has the same file length as the genuine verification software, with the only difference being it always proclaims a "match" regardless of the integrity of the image being verified. With the software freely and widely available this is not a large risk, as additional copies can be easily obtained from multiple sources and a best 2-out-of-3 scheme can be employed. When the stakes are high and it is extremely important that the verification software be known to be genuine, an independent certification authority or the manufacturer could then be called in to provide their

own copies of the software and their own lists of public keys at the time of verification.

The algorithms and private key necessary for encrypting the additional digital signature file from within the camera are to be embedded inside a new breed of secure microprocessors whose ROM contents cannot be observed outside of the factory. Because the private key used for encryption is hard-coded into this chip by the manufacturer (who must then ensure the private key remains secret), credibility of the camera's output becomes an extension of that of the manufacturer; a digital signature from the camera can be considered to be just as reliable and secure as if the signature had been generated by the manufacturer.³

Each camera should possess its own unique pair of private and public keys, with the private key etched into the camera's secure microcontroller and the public key stored in three places: in a public key list kept by the manufacturer, on the camera body itself (which can then also double as the camera's serial

number), and in the colorful border that contains more data about the captured image (see "A Special Border" section below for more details on this idea). Assigning unique keys to each camera has the benefit of avoiding instant obsolescence which would occur if only one private key were used for all cameras, and that key were

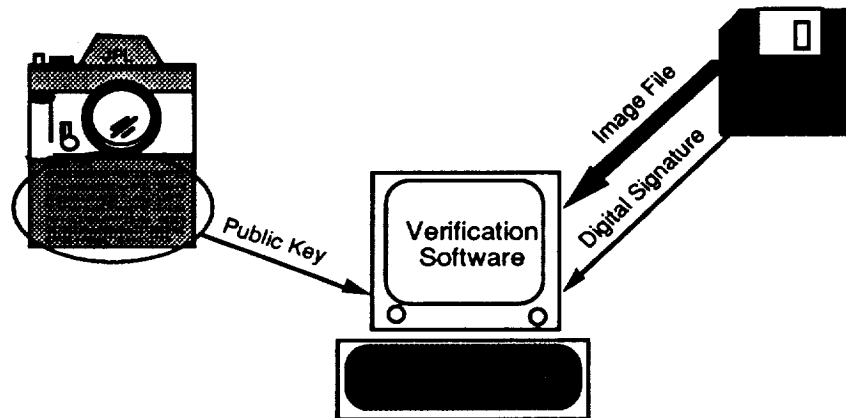


Figure 5: To authenticate the image, public domain verification software is run on a standard computer platform. The program takes as input the image file in question, the digital signature, and the camera's serial number (which doubles as its public key).

to be compromised. An even higher level of security

³ Any company involved with the development of a Trustworthy Digital Camera would have to address the issue of liability, for if the security of the private key were ever to be compromised (for example by a disgruntled employee who steals a private key and uses it to generate false authenticatable images), the lawsuits brought on as the result of a false positive would necessitate significant insurance coverage.

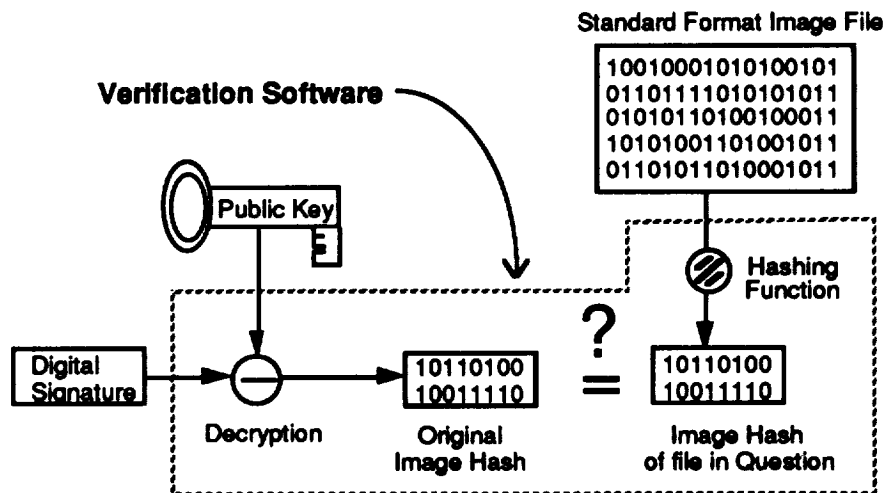


Figure 8: The verification software computes its own hash of the image in question, and compares it to the original hash which has been decrypted using the public key. If the image in question has not been manipulated, the decrypted digital signature and the program's own hashing function will match, resulting in an authentication. If even a single bit is different, the two hashes will not even closely match, yielding an authentication failure.

would occur if the manufacturer were to destroy all records of the private key once the camera is produced. (At that point the private key is no longer needed.) This would eliminate the possibility of compromise via industrial espionage or theft.

Finally, regular and free distribution of all valid public keys is desirable to defeat a counterfeiter who has learned of the encryption algorithm employed and has written a program to produce digital signatures based on his own private key. Decoding these forgeries would require the use of a public key not generated by the manufacturer. Freely distributing updated public key lists would make it easy to identify and thwart such attempts.

Uses

The single most obvious use of a trustworthy camera would be in situations where proof of image authenticity is necessary; such as for legal evidence or insurance claims. The inevitable transition to digital cameras and electronically-transmitted images will also make it more difficult for the professional photographer to protect his or her image copyright, since with electronic cameras there is no "original" to control, and works stored in computer format tend to proliferate faster and with less control from the author than the traditional distribution method (which places image control in the hands of whoever holds the original negative or transparency). Just as it is common practice today to obtain model releases for any published picture containing a recognizable face, it is foreseeable that no electronic

image in the future will be published without first having authenticated the image using the digital signature of the camera which was or is registered to the photographer.

This technique need not be limited to still digital images. Because digital signatures can be used to verify any block of digital data, it can also be engineered into digital video cameras and digital audio tape recorders. In both these devices, a digital signature can be generated and recorded onto the medium each time the recording process stops or pauses; this way each sound byte or video "take" is hashed, encoded and written at the time it's created.

A Special Border

Since the proposed camera is being initially targeted towards legal authentication, a few additional features can be implemented to better serve this use. A brightly-colored border could automatically be generated as part of each captured image file. Within the border would appear textual information about the image: the date and time it was taken, the ambient light level seen by the camera at the time of exposure, the original color temperature of the scene, the software version of the camera's firmware, the camera's serial number, the focusing distance of the lens at the time of exposure, a unique sequence number, and (when the technology allows for a Global Positioning System (GPS) receiver to be built into the camera) the geographical coordinates of the camera, indicating where in the world you were when the picture was taken. The ambient light level and

color temperature readings would be useful for getting a feel for exactly what the scene was like at the time of exposure; something a sensitive optical element might inadvertently hide via automatic exposure and color correction. The lens' focused distance is there to help detect potential abuse of the trustworthy camera: taking close-up pictures of a modified photo and trying to pass it off as an unaltered original. Since all these textual data in the colored border are part of the authenticated image file, their credibility is also upheld when authenticated by the verification software.

The accuracy of the date and time information would again be the responsibility of the secure microprocessor; in addition to being able to keep its programming a secret, it also would have a lithium battery powering a system clock that was set to Universal (Greenwich Mean) Time at the time of manufacture. If the timer circuit ever fails or is tampered with, the system will be programmed to fill the time and date fields with XXXX's, eliminating the chance of a random time stamp being mistaken for the actual time.

Higher Level of Security

Although the proposed Trustworthy Digital Camera offers a satisfactory level of security, nevertheless there still exists a small possibility that a determined saboteur will be able to crack the camera's private key given an extended amount of time. (No cryptographic scheme will protect your data forever; given sufficient time, advancements in code breaking or improved computer horsepower will be enough to render any given level of cryptographic protection obsolete.) If the discovered private key were then to be published, it would allow an individual to generate authentic-looking digital signatures on altered image files, essentially undermining the credibility offered by the compromised camera. (The security level of other cameras in use, and of images taken with those cameras, will still remain high.)

Because of this risk, it would be wise for a manufacturer of such cameras to regularly upgrade and enhance the sophistication of the encryption implementation as newer camera models are introduced, typically using longer encryption/decryption key lengths and improved encryption/decryption algorithms. It is expected that evolving verification software (the public domain software component of this authentication scheme which is freely distributed) will then be designed to recognize, identify and authenticate all previous versions.

Because the encryption details must necessarily be changed often (depending on the technological capabilities of the day), no single image format, key

length or digital signature algorithm is being specified in this disclosure⁴.

Conclusion

The Trustworthy Digital Camera is an application of existing technology toward the solution of an ever-more-troubling social problem, the eroding credibility of the photographic image. Although it will always be possible to lie with a photograph (using such time-honored techniques as false perspective and misleading captions), this proposed device will prevent the explosion of very capable personal computers from driving up the incidence of doctored photographs being passed off as truth.

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Bibliography

- [1] "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithm", IEEE Transactions on Information Theory, Vol. IT-31 no. 4, (July 1985), Taher El Gamal, pp. 473-81.
- [2] "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems", Communications of the ACM, Vol. 21 no. 2 (February, 1978), R.L. Rivest, A. Shamir, L. Adleman, pp. 120-26.
- [3] "New Directions in Cryptography", IEEE Transactions on Information Theory, Vol. IT-22 no. 6 (November, 1976), Whitfield Diffie and Martin E. Hellman, pp. 644-54.
- [4] "The Proposal for a U.S. Standard for Digital Signature Encoding", IEEE Spectrum, August, 1992, Dennis K. Branstad, pp. 30.
- [5] "8051 Tackles Secure Smart-card Applications", EDN, December 10, 1992, pp. 46-48.
- [6] "Photography by the Numbers", BYTE, January 1993, Howard Eglowstein, pp. 241-244.

⁴Although the National Institute of Standards and Technology's (NIST's) proposed Digital Signature Standard (DSS) was in mind when this method was conceived, adherence to it, its algorithms, formats, or royalties is not required for implementation.

omit

VIRTUAL REALITY/ SIMULATION

Virtual Reality In Medical Education And Assessment

Laurie A. Sprague, Brad Bell, Tim Sullivan, Mark Voss
LinCom Corporation, 1020 Bay Area, Houston, TX 77058

Dr. Andrew F. Payer, Ph. D.
University of Texas Medical Branch, 200 University Blvd, Galveston, TX 77555-0843

Stewart Michael Goza -National Aeronautics and Space Administration
Johnson Space Center, Houston TX 77058 Mail Code ER6

ABSTRACT

The NASA Johnson Space Center (JSC)/LinCom Corporation, the University of Texas Medical Branch at Galveston (UTMB), and the Galveston Independent School District (GISD) have teamed up to develop a Virtual Visual Environment Display (VIVED) that provides a unique educational experience using Virtual Reality (VR) technologies. The VIVED end product will be a self-contained educational experience allowing students a new method of learning as they interact with the subject matter through VR. This type of interface is intuitive and utilizes spatial and psychomotor abilities which are now constrained or reduced by the current 2 dimensional (2D) terminals and keyboards (2). The perpetual challenge to educators remains the identification and development of methodologies which conform the learners abilities and preferences. The unique aspects of VR provide us with an opportunity to explore a new educational experience.

Endowing medical students with an understanding of the human body poses some difficult challenges. One of the most difficult is to convey the three dimensional (3D) nature of anatomical structures. The ideal environment for addressing this problem would be one that allows students to become small enough to enter the body and travel through it - much like a person walks through a building. By using VR technology, this effect can be achieved; when VR is combined with multi-media technologies, the effect can be spectacular.

INTRODUCTION

Medical students, interns, residents and physicians spend a significant amount of their time identifying and organizing data and information based on their educational programs or interactions with patients. It is important that they effectively organize this data and identify any issues that they need to learn about. Many of these "learning issues" can be accessed through the computer. The use of computers to process medical information, known as "Medical Informatics," is becoming more frequently used. This technology has advanced to 3D renderings of human anatomy on a 2D computer screen. Even with this technology and other classic methods, such as dissection of cadavers, there are times when it would be helpful for the user to be able to place himself into the anatomical environment to better understand the true 3D relationships that exist in that particular region of the body. This leads to the need for VR experiences in anatomical imaging. Studying images on a 2D computer screen can be compared to looking at fish through a glass bottom boat (1, 3), whereas VR allows one to put on the scuba equipment and enter the water, interacting with the surroundings without getting wet. This type of learning experience would be an excellent addition to the currently available methods.

With inevitable improvements in imaging for VR and in digital gloves, it is not difficult to visualize future applications for this technology. The practice of general medicine can be compared to activities of an airline pilot, whereby each deals with many routine landings and takeoffs, both also have to be prepared for emergency situations. We have the means to simulate the environment for the pilot allowing him to develop specific skills in routine and emergency situations. Imagine a physician putting on a VR helmet and digital gloves for a clinical

simulation. When the computer is activated, the physician sees the operating room, equipment, staff and patient. He/she holds out a hand, feels a scalpel being placed in it and begins a surgical procedure allowing him to practice the procedure and evaluate any possible problems before opening up the patient.

Fidelity of the images produced is of utmost importance when applied to medical education. Magnetic Resonance Imaging (MRI), Computerized Axial Tomography (CT) scans and <1 millimeter cadaver sections all provide acceptable levels of image quality. The National Library of Medicine is using these methods of data acquisition to build a nationally accessible database of information on human anatomy. This project, known as the Visible Human Project, will provide an invaluable source of data for VR work in medical education.

NASA JSC has used VR to expand its training capabilities. Specifically, by providing a means to incorporate extravehicular activities (EVA) with remote manipulator system (RMS) task training. This same VR technology is being transferred by way of the VIVED project for use in medical education. This is the subject of this paper.

BACKGROUND

In 1990, Dr. Andrew F. Payer was invited to experience a VR prototype for astronaut training that utilized wire frame models of the shuttle and space station. He was placed in the VR module and allowed to translate around objects under his own control. After the experience, he was asked if he thought there were any possible applications to medical education. This was the beginning of an exciting project called the Virtual Visual Environment Display (VIVED) project.

Discussions with potential commercialization partners identified the need for the project to have applications broader than just the medical school environment (there are only 137 medical schools in the US). It was decided that a joint plan be identified that would allow the VR prototypes to be tested for effectiveness in educational arenas including elementary schools, high schools, colleges and professional schools. Thus a Memorandum of Understanding was developed between NASA JSC, UTMB, and GISD so that prototypes could be tested in a variety of educational environments.

The human skull was chosen as an initial specimen due to readily available methodology, namely CT scans, of obtaining imaging data for bony structures. A second model of a heart specimen was created using MRI data.

Early models generated for VR did not meet the quality requirements of the project. After using newer CT scanners of higher resolution, and improvements in VR software, the skull image now nears the quality of an actual human skull. Because of the amount of image data, interactive fly-throughs into the skull are not currently possible. However, prescribed fly-throughs are being generated in digital and video formats to utilize multimedia technology.

The short term goal of VIVED is to integrate prescribed fly-throughs of the skull with other interactive multi-media (audio, video, etc.) capabilities.

METHODS: CREATING STEREO SEQUENCES OF THE HUMAN SKULL

File Conversion and Data Preparation

Scans of a human skull were performed resulting in a data set consisting of over 120 slices through the skull and 60 slices through the mandible (jaw). The first attempts used 2.5 mm slice thickness and a low resolution scanner resulting in poor image quality. The last scan used a newer high resolution scanner and a slice thickness of 1.5 mm improving the quality tremendously (see Figures 1 and 2).

The data files obtained from the scan were transferred to JSC IGOAL (Integrated Graphics, Operations, and Analysis Laboratory) for further processing. The skull was held in place during the CT scan by a foam band thus creating extraneous data. The scans were then cropped to eliminate as much extraneous data as possible without losing any critical information. The final stage of data file preparation required using a tool developed in the IGOAL called "ctimager" to remove unwanted noise and extraneous data from each slice.

The data files obtained from the scan were transferred to JSC IGOAL (Integrated Graphics, Operations, and Analysis Laboratory) for further processing. The skull was held in place during the CT scan by a foam band thus creating extraneous data. The scans were then cropped to eliminate as much extraneous data as possible without losing any critical information. The final stage of data file preparation required using a tool developed in the IGOAL called "ctimager" to remove unwanted noise and extraneous data from each slice.



Figure 1 First Skull attempts using CT data



Figure 2 Current 3D Skull Model

A MRI scan of a human heart was performed resulting in a data set consisting of 200 slices. In order to use the image tools developed to process the skull data, the data for the heart was converted to the same format as the skull data. Ctimager was then modified to automatically remove the noise from each scan.

Data Filtering And Conversion Of Volume Data To Polygonal Data

A tool called "disply" was developed in the IGOAL to convert volume data into a form that could be displayed directly by the computer. Disply used multiple filtering algorithms to prepare the CT and MRI data for conversion to polygonal form. Anatomical models are generated based on the marching cubes algorithm developed by W. E. Lorensen and H. E. Cline. This technique generates surfaces based on the density of the imaging data. It generates a polygonal surface based on a "logical" cube created from eight data points on two adjacent slices of data. The algorithm determines how the surface of a selected density intersects this cube. A table lookup is performed to determine which triangles describe the surface. The normals are then determined using gradients computed directly from volume data.

The raw data set contains noise from the CT or MRI scanning process, therefore data is filtered before generating the polygonal model. The filtering process typically consists of thresholding the data to eliminate most of the noise. A low pass filter is used to minimize the high-frequency noise that would produce an irregular bumpy surface when input to the algorithm. This filtering process produces a relatively smooth surface that approximates the scanned specimen and reduces the number of polygons generated due to noise. A unique filter was created for the heart data which only smoothed the data between scans, no other filtering was needed (see Figure 3). Due to the large number of slices in both heart and skull data sets, several models were made, each of which represented a small number of slices. To improve display performance a meshing algorithm was developed (meshit), that converted the raw collection of triangles into efficient strips. The triangle strips averaged over 100 triangles in size.

Generating Stereo Images

After the models were made, stereo sequences were rendered. To generate the sequences, a tool developed by IGOAL called OOM (Object Orientation Manipulator) was used. OOM (available through COSMIC) rendered each frame to disk. The images used red and blue color separation for representing stereo images. Once the sequence was recorded to disk it was converted frame by frame to Macintosh ".pict" format and transferred to a Mac (full color image sequences were also transferred to the Mac for non-stereo viewing).

Macintosh - Stereo Images and Multi-Media

The Macintosh was selected because it is a relatively affordable platform and has a wide base already installed in the school systems. It is a leading engine of desktop multi-media and has a wide array of software and hardware available for this task, though more is needed.

Once on the Macintosh, the images were edited to produce the desired effect, such as digitized cadaver overlays or text inserts describing what is being viewed. By using Apple's QuickTime extension, the images were then converted into QuickTime movies for animation on the Mac. The movies can also be edited on the Mac. To be able to view the stereo effects, the user must wear red/blue 3D glasses.

A Hypercard interface (a simple tool for creating prototypes) is being created to house the Educational Experience on the Mac. Other hypertext programs have been investigated and will eventually replace the Hypercard interface. Many forms of multi-media will be considered and then incorporated into the finished product including, but not limited to, a Virtual Reality head mounted display (HMD) or boom system, CD ROM, laser disc, audio, video, digital imagery, and 3D.

The VIVED application will contain lectures, supplemental text, graphics, digital movies, notes, pop quizzes, references, exams, Virtual Reality experiences of anatomy, student comments, bookmarks, qualitative analysis of student performance, summations, and program evaluations.

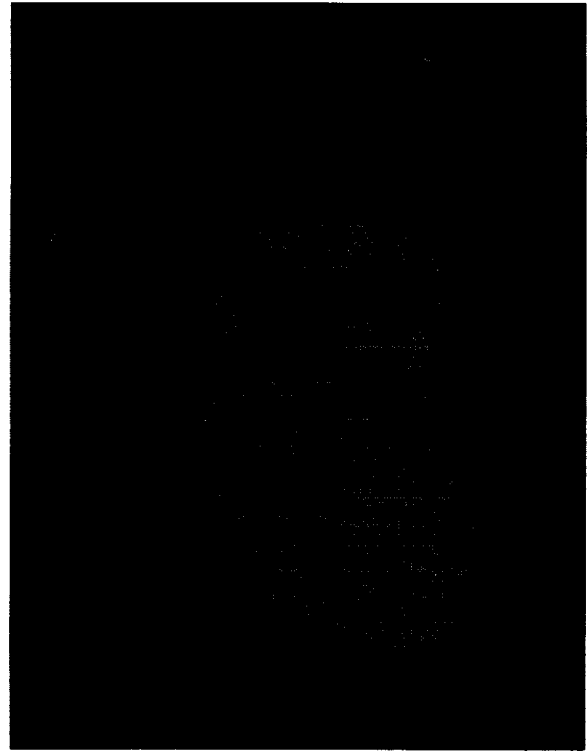


Figure 3 3D Heart Model from MRI data

CONCLUSIONS

The VIVED project is at a point where CT scanned medical images of bone (e.g., the skull) can be generated into high quality VR imaging for prescribed fly-throughs on the Macintosh computer using a HMD or boom system. The project team is in the process of working on a heart VR model that has been generated from MRI data. Preliminary results indicate that a high resolution model can be developed using this type of imaging data. The project has been able to maintain its goal of high quality of VR imaging. This has led to some problems because of the amount of data the computer needs to deal with even during frame-by-frame sequencing of the prescribed fly-throughs. Alternative hardware and software options are being explored to deal with this problem.

Another problem has been the current status of technology for the display systems for HMDs. The LCD displays do not have the resolution needed to maintain a high quality VR experience. The CRT displays are reaching the resolution needed for the project, but the cost is prohibitive for multiple education platforms.

The next goal for the project include improving the software and hardware for generating VR images, developing prescribed fly-throughs and incorporating multi-media into the VR fly-throughs. Other anatomical imaging data will be obtained from CT scans, MRI and Cadavers to develop VR imaging of anatomical regions that contain different tissues with different data densities

REFERENCES

1. Brooks, F. P. 1988. "Grasping Reality Through Illusion—Interactive Graphics Serving Science." *Proceedings AMC SIGCHI*.
2. Fuchs, H., Levoy, M., and Pizer, S. M.. 1989. "Interactive Visualization of 3D Medical Data." *IEEE Computer*, August.
3. Furness, T. 1987. "Designing in Virtual Space." Chapter in W.B. Rouse and K.R. Boff, eds., *System Design: Behavioral Perspectives on Designers, Tools, and Organizations*. North Holland.

TECHNOLOGY TRANSFER OF OPERATOR-IN-THE-LOOP SIMULATION

K. H. Yae

H. C. Lin

T. C. Lin

Simulation and Design Optimization of Mechanical Systems
Department of Mechanical Engineering

The University of Iowa

Iowa City, Iowa 52242

and

H. P. Frisch

NASA Goddard Space Flight Center

Greenbelt, MD 20771

October 7, 1993

Abstract

The technology developed for operator-in-the-loop simulation in space teleoperation has been applied to the Caterpillar's backhoe and wheel loader, and off-highway truck. On a SGI workstation, the simulation integrates computer modeling of kinematics and dynamics, real-time computation and visualization, and an interface with the operator through the operator's console. The console is interfaced with the workstation through an IBM-PC in which the operator's commands were digitized and sent through a RS 232 serial port. The simulation gave visual feedback adequate for the operator in the loop, with the camera's field of vision projected on a large screen in multiple view windows. The view control can emulate either stationary or moving cameras. This simulator created an innovative engineering design environment by integrating computer software and hardware with the human operator's interactions. The backhoe simulation has been adopted by Caterpillar in building a virtual reality tool for backhoe design.

Key Words: Teleoperation, redundant manipulator, simulation, recursive dynamics

1. Introduction

Teleoperation [1,2] involves a human operator, a hand controller, and a manipulator, in which the manipulator under the operator's supervision completes tasks ranging from simple trajectory following to pick-and-put operation. The operator's interaction with the manipulator becomes a key issue, because of the intrinsic difference between a human operator and a robotic manipulator; that is, a human operator works more efficiently in Cartesian space [3] whereas most manipulators are designed with joint servo control.

For off-line trajectory analysis, the inverse position and orientation problem can be solved iteratively using inverse velocity/angular velocity solutions [4]. In teleoperation, however, the desired configuration of the end-effector is known only after the operator has commanded through the mini-master. This means that both Jacobian construction and inverse kinematic analysis must be completed on-line in real time. Unlike the Jacobian defined for the pre-determined end-effector's trajectory, the constraint Jacobian can be constructed on-line as the operator controls the manipulator's end-effector through the mini-master.

The constraint Jacobian is derived from the six constraints [5] that are imposed between the current and the desired end-effector's Cartesian position and orientation. The desired Cartesian position and orientation of the end-effector is viewed as the target that the current position and orientation will eventually have to assume. Constraining the two sets of position and orientation yields six constraints from which the constraint Jacobian is constructed [6]. This Jacobian and its pseudoinverse are used in iterations to yield the joint command angles necessary for the joint controllers. The procedure is illustrated with a Kraft 6-d.o.f. mini-master and a 7-d.o.f. redundant manipulator [7] and also with an operator's console and a backhoe.

Since the manipulator's dynamics involves kinematic redundancy and intermittent kinematic loop closure, the dynamic model is constructed in the recursive Newton-Euler dynamic formulation adapted for high-speed simulation of general constrained mechanisms [8,9,10,11,12]. According to d'Alembert's principle, this formulation starts out with the virtual work of all the links and the cut joint, where the cut joint is expressed in the Lagrange multiplier

and the Cartesian coordinates that include both independent and dependent coordinates. The dependent coordinates are then replaced with independent joint coordinates. The replacement is recursive and systematic and it has been automated [8,10]. Consequently, this method facilitates modeling by allowing formulation in Cartesian coordinates, expresses the model in terms of independent joint variables, and improves computational efficiency. As a result, this formulation can efficiently generate the equations of motion of a constrained and/or unconstrained multibody system for a single processor computer [8] and for a multi-processor computer [10]. The manipulator's dynamic equations of motion are then combined with the control system. This model of dynamics and control is then teleoperated by an operator through the mini-master.

2. Dynamics and Control of Manipulators

Being driven interactively by the operator-in-the-loop, the simulation requires that dynamics and control computation and graphics rendering be fast enough to provide the operator with adequate visual cues without too much delay. For high-speed computation, a recursive Newton-Euler dynamic formulation has been used to model the manipulator and vehicle dynamics. The recursive formulation derived in [8,9,10,11,12] facilitates modeling a general constrained mechanical system and improves computational efficiency.

2.1. Recursive dynamic formulation in spatial vectors

The basic procedure for deriving the equations of motion based on the Newton-Euler formulation can be understood if we view it as a way to calculate the joint driving force and torque necessary to realize a given trajectory of the joint coordinates.

For illustration, let us consider a tree-like open chain of rigid bodies with their joint constraints known. Suppose that the current values of joint displacements, velocities, and accelerations are known and that the force and torque exerted on the end link by the environment and a grasped object are given. First, we calculate the angular velocity, the angular acceleration, the linear velocity, and the linear acceleration of a link with respect to the reference frame. Second, using Newton's and Euler's equations, we calculate the force and torque that must be applied to the center of mass of a link to realize such a motion. Third, we calculate the force and torque that must be applied at a joint to produce the corresponding force and torque, starting from the end link and moving inward to the base, with the given values of the force and torque at the end link. Finally, we calculate the joint driving force (and/or torque) at each joint [13].

The kinematics of two contiguous bodies is defined using the spatial velocity vector [14], also called the velocity state vector in [8]. The use of spatial vectors simplifies the derivation and improves computational efficiency in dynamic simulation. Each link has two body-reference frames located not necessarily at the center of gravity but at two joints, proximal and distal. Each joint connects a proximal (inboard) link and a distal (outboard) link. Starting from the base body, the recursive kinematics is developed from a proximal link to a distal link. After the kinematics has been defined, starting from a tree-end body, the recursive dynamics is developed in the opposite direction, i.e., from a distal link to a proximal link.

The spatial velocity vector represents the velocity of a point that is moving with the body but is instantaneously coincident with the origin of the global inertial coordinate [14]. An interesting property of this vector is that it remains invariant no matter where the body reference frame is located within the body [10]. The spatial velocity vector is then the sum of that of its inboard body's spatial velocity vector and the angular velocity of the joint.

The virtual work of the manipulator is first expressed in the variations of spatial vectors [8,10,12,15]. When there exists a kinematic loop in the system, the cut-joint technique [16] is used. This equation is then converted to the recursive equations of motion in joint variables. The conversion proceeds from the outmost body to the base body.

2.2. The power train model of 950F wheel loader

The wheel loader dynamics requires manipulator dynamics for the loader and vehicle dynamics for the tractor. The engine's driving torque is transmitted through a gear train to the four tires. The wheel loader is then driven by the tire-ground reaction forces.

The gear train consists of torque converter, clutch, brake, and final reduction gears. The computation of the power train starts from the torque converter model. The input and output torque of the torque converter are determined first

based on the input speed (engine RPM) and the output speed (clutch speed or vehicle speed). The input torque to the torque converter is viewed as the engine load. The converter output torque is the input to the rest of the power train. This output torque is transformed into wheel driving torque through the gear train. Once the wheel velocity is computed, the driving force and torque acting on the chassis is then computed through the tire-ground interaction. During the simulation, the wheel speed is updated by integrating the accelerations of the engine and the wheels (tires).

Engine model: A simplified empirical engine model is developed for real-time simulation. The engine's torque generation is determined by the engine speed and the throttle. The net torque applied to the equation of motion is equal to the sum of the engine torque and the torque reflected by the torque converter. This net torque yields the engine acceleration, which is then integrated to give the speed for the next time step. The generation of engine torque can be characterized by three curves: power, friction, and governor curve as shown in Figure 1. The output torque curve consists of three segments that are determined by the throttle percentage. Once the torque curve is known, the actual output torque is determined by the engine's running speed. These curves are constructed based on the experimental data.

Tire and brake model: For real-time simulation, no differential gear is modeled in dynamic analysis. The output torque from the transmission is equally distributed to the four wheels. In addition, each wheel can rotate freely with no coupling with other wheels. A simplified model is developed for computing the tire-ground reaction forces, based on two assumptions: 1) the ground is flat and 2) the tire contacts the ground vertically at a point with an effective radius (deformed radius). While a tire is rolling, the translational direction of the chassis may not align with the tire's longitudinal direction. The angle between the translational and the longitudinal direction is called the slip angle. The slip index that represents the longitudinal deformation of a tire is used to compute the tire-ground interaction force in traction (longitudinal) and cornering (lateral) components. These two components give the driving force to the wheel loader.

2.3. Actuators and Control Systems

The control and actuation system generates the driving force and torque that is fed into the dynamic model. As for the telerobotic manipulator, the controller has been modified for real-time simulation and validated with the original controller's model through actual experiments conducted in the Robotics Laboratory at NASA Goddard Space Flight Center. Finally the simplified controller and the dynamics are put together. The nonlinear dynamic model feeds back joint position and velocity to the joint controller, and the joint controller generates driving torque that drives the manipulator.

In the wheel loader shown in Figure 1, four hydraulic systems drive the swing, plate, and Z-bar linkage. Each system includes a spool valve and a cylinder. The hydraulic actuator is modeled as a five-port system, with the assumptions of constant supply and drain pressure, compressible fluid, and zero-lapped spool valves.

A mathematical model of the hydraulic actuator has been first developed and then modified to include static friction force and damping force. The model consists of a set of differential equations that relate the time derivatives of two pressures in one cylinder with the flow rates and the piston velocity. The cylinder pressures are then obtained by integrating the differential equations. The pressure difference yields actuating force in a cylinder.

3. Interface with the Kraft mini-master and an operator's console

A situation unique to teleoperation is that the target position and orientation is entered on-line by the operator. Consequently, the control algorithm must first translate incremental Cartesian coordinates into incremental joint angle change.

As for the wheel loader, the operator's inputs are accelerator pedal, brakes, gear shift, and bucket lift and tilt. Some of them are interpreted as spool valve displacements. They all are merged into dynamics, control, and hydraulic actuator models without any need for inverse dynamics.

3.1 Cartesian Space Control of a Telerobotic Manipulator with the Kraft Mini-Master

The telerobotic manipulator used in this research is a 7 d.o.f. device with shoulder roll and pitch, elbow roll and pitch, wrist roll and pitch, and toolplate roll. The mini-master in Figure 3 is a 6 d.o.f. force-reflective device

kinematically similar to a human arm, with shoulder yaw and pitch, elbow pitch, and wrist yaw, pitch, and roll. Even with one joint of the 7 d.o.f. manipulator being locked, the one-to-one mapping of joint angles between these two simply confuses the operator. It is, therefore, necessary to translate the mini-master's joint angle readings into the joint angle commands to the manipulator's joint controllers.

For the interface with the mini-master, we found it easier to first build the mini-master's kinematic model and expand the mini-master's workspace to the manipulator's workspace as closely as possible. The mini-master's kinematic model converts three joint angles (shoulder yaw and pitch and elbow pitch) into the hand grip's Cartesian position in the mini-master's workspace; then, this Cartesian position is transformed into the Cartesian position of the manipulator's end-effector, which is then converted into four (lower) joint angles of the manipulator through a generalized inverse method. Similarly for orientation, the other three (upper) joint angles of the mini-master (wrist yaw, pitch, and roll) are mapped onto manipulator's three joint angles (wrist roll and pitch and toolplate roll).

The control algorithm must first translate incremental Cartesian coordinates into incremental joint angle change [4], and then the incremental joint angle change into joint torque. This translation requires inverse kinematic analysis of a manipulator as described in the following.

The end-effector's Cartesian position and orientation is defined by the joint angles through a kinematic relation. For known Cartesian coordinates, instead of seeking an explicit expression for the joint angles, we view the kinematic relation as constraint equations. From their variations the constraint Jacobian is defined.

When the telerobotic manipulator is kinematically redundant, a single Cartesian position of the end-effector may correspond to multiple sets of associated joint angles. If an additional condition is added, such as a minimization of the Euclidean norm of angular velocity [4,7,17], a unique set of joint angles can be identified. This method is called the pseudoinverse method, or Moore-Penrose generalized inverse method [18]. Although using the pseudoinverse for the derivatives does not yield an inverse function between the variables themselves [19], the pseudoinverse method turns out to be useful for real-time on-line computation as is required in teleoperation.

Teleoperation is initiated by the operator's command input through the mini-master. The operator's input motion is divided into a series of increments in Cartesian space. These Cartesian increments are transformed into incremental angles by the Jacobian. The incremental joint angles are then input to the joint servo controllers, which in turn produce necessary joint torque. The resulting position and orientation is displayed on a graphics workstation to provide the operator with visual feedback.

The Kraft mini-master (Figure 3) is kinematically similar to a human arm so that the human operator can "wear" it comfortably for use in Cartesian space with minimal cognitive learning [20]. The manipulator of seven degrees of freedom, on the other hand, is joint-controlled. Therefore, the task is to relate the mini-master's six joint angles to the seven joint angles of the manipulator in the way that the manipulator's end-effector follows the mini-master's grip.

The host computer, an IRIS 4D/320 VGX, provides enough computational power for high-speed simulation of the dynamic and control system, as well as high-speed graphics rendering and data communication between the host computer and the mini-master. Through a RS 422 serial port, the Kraft mini-master is capable of sending out the readings of its own joint angles and receiving feedback joint torque for tactile feedback. When the sensed joint angles reach the host computer, they are converted into the joint input commands. The feedback joint torque received can represent a payload or reaction force and torque for the end-effector in contact with the environment.

3.2. Interface between operator console and simulation

The setup of simulation is shown in Figures 4 and 5. During the simulation the operator receives visual feedback and controls the wheel loader through the operator's console. The console generates three types of output: the analog signal from the rotational potentiometers, digital signal through digital I/O port, and the encoder output signal. All these inputs are consolidated on a PC and sent to the host computer, IRIS 4D/320 VGX, through RS 232 ports, as shown in Figure 6. The PC digitizes analog signal, receives digital data through digital I/O, counts pulses of the encoder output, and transmits data through RS 232 serial ports to the host computer. In the host computer, the dynamics and the graphics program run concurrently and share data through shared-memory interface.

An incremental encoder is used for detecting the operator's steering inputs. The spool displacement of hydraulically-actuated steering is proportional to the turning rate of the steering wheel, instead of its rotational displacement. Thus a high-resolution encoder is used to give high-resolution rotational displacement from which a turning rate is obtained. The encoder generates 1200 pulses per revolution. The encoder has two outputs, between which there exists a 90-degree phase shift, that can tell the direction of rotation and can also quadruple the resolution. Thus, the total resolution of the encoder is 4800 pulses per revolution.

Five rotational potentiometers are used to detect the gas pedal, brake, clutch, lift, and tilt levels. The output from these sensors is digitized through an 8-channel 12-bit A/D converter. The output from the gear switching box is a 7-bit digital signal and is decoded through a digital I/O port.

4. Applications to Telerobotic Simulation and Wheel Loader Operation

Now that the dynamics, control, and actuator models are ready and the interface with the Kraft mini-master and the operator's console have been defined, they must all be integrated with computer graphics. When the computer graphics give visual feedback to the operator, its fidelity becomes important. The graphics should include a general perspective view for depth perception, as well as proper rendering attributes such as color, surface texture, shading, and lighting. In addition, the graphics rendering speed should be fast enough not to show jerky motion.

For the interactive simulator, the Visualization of Dynamic Systems (VDS) [21] was modified [22,23]. This graphics software needs three data files to display a manipulator in motion. Two of them, created before the simulation, define the geometry and the rendering attributes. The geometry of each link can be generated by MOVIE.BYU [24] or other compatible software. The third file (or data stream), created during the simulation and updated at each frame, contains the latest position and orientation data computed from dynamics and control.

In Figure 7 three views from the on-board cameras on the tele-robot and one view from a camera on the space shuttle (lower right in Figure 7). The first view point is located right on the RMS end-effector and looks straightforward (upper right window in Figure 7), where the ball indicates the location of the telerobot's end-effector. The second one is located 2.5 ft behind, 1 ft left, and 1 ft up from the first one (upper left window in Figure 7). The third one is 3 ft away from the first one to the right and looks toward the first one (lower left window in Figure 7). The RMS is operated by a pair of joysticks and the tele-robot is operated by the force-reflective mini-master, as shown in Figure 3.

Figure 8 shows the overall view of the wheel loader simulation. When the simulation starts, the position analysis computes the position and orientation of each object based on the system properties and the initial conditions. The position and orientation is then sent to the computer graphics package, and combined with the prepared geometric model data, for animation. With a minor modification, the wheel loader simulation is extended to the simulation of an off-highway truck.

On a two-processor IRIS workstation, the dynamic and control analysis is executed on one processor and graphics rendering on the other. Based on the position information, the joint velocities are computed for use in control analysis, which computes the control driving torque. The joint torque is then sent to acceleration analysis. Next, the computed acceleration is integrated to update velocity and position. This update is then sent back to position analysis for the next time step. In Figure 3, the Kraft mini-master is shown in the foreground and the simulated manipulator in the background on the graphics screen. In Figure 9, the operator's view is displayed with two side mirrors' view, as seen from the off-highway truck's console in Figure 5.

5. Conclusion

The dynamics and control model of a manipulator has been teleoperated through a 6 d.o.f. mini-master. A high-speed operator-machine interactive simulation includes the manipulator's recursive dynamics, control, high-speed graphics, and interface with the mini-master and the operator's console. This technology developed for operator-in-the-loop simulation in space teleoperation has been applied to the Caterpillar's backhoe, wheel loader, and off-highway truck. On a SGI workstation, the simulation integrates computer modeling of kinematics and dynamics, real-time computation and visualization, and an interface with the operator through the operator's console. The console is interfaced with the workstation through an IBM-PC in which the operator's commands were digitized and sent through a RS 232 serial port. The simulation gave visual feedback adequate for the operator in the loop, with the camera's field of vision projected on a large screen in multiple view windows. This simulator created an

innovative engineering design environment by integrating computer software and hardware with the human operator's interactions.

Acknowledgment

Research has been supported by NSF-Army-NASA Industry/University Cooperative Research Center for Simulation and Design Optimization of Mechanical Systems in The University of Iowa.

- [1] J. Vertut and P. Coiffet. Teleoperations and robotics: evolution and development. In Robot Technology. Prentice-Hall, 1986.
- [2] T. B. Sheridan. Telerobotics. *Automatica*, 25(4):487-507, 1984.
- [3] L. Stark. Telerobotics: Display, control, and communication problem. *IEEE J. Robotics and Automation*, RA-3(1):67-75, 1987.
- [4] C. W. Wampler. Manipulator inversekinematic solutions based on vector formulations and damped least-squares methods. *IEEE Trans. Systems, Man, and Cybernetics*, pages 93-101, 1986.
- [5] E. J. Haug. Computer Aided Kinematics and Dynamics of Mechanical Systems, Volume I: Basic Methods. Allyn Bacon, 1989.
- [6] S. T. P. Chern, K. H. Yae, and T. C. Lin. An application of the constraint Jacobian to teleoperation of a redundant manipulator. *Int. J. Robotics Automation*, 1993. to appear.
- [7] D. N. Nenchev. Redundancy resolution through local optimization: A review. *J. Robotic Systems*, 6(6):769-798, 1989.
- [8] D. S. Bae and E. J. Haug. A recursive formulation for constrained mechanical system dynamics: Part I -open loop systems, Part II-closed loop system. *Mechanics of Structures and Machines*, 15(3,4):359-382 and 481-506, 1987.
- [9] D. S. Bae, R. S. Hwang, and E. J. Haug. A recursive formulation for real-time dynamic simulation of mechanical systems. *ASME J. Mechanical Design*, 113:158-166, 1991.
- [10] F. F. Tsai and E. J. Haug. Real-time multibody system dynamic simulation Part I: A modified recursive formulation and topological analysis, Part II: A parallel algorithm and numerical results. *Mechanics of Structures and Machines*, 19:99-127 and 129-162, 1991.
- [11] T. C. Lin and K. H. Yae. The effects of harmonic drive gears on robotic dynamics. In *ASME 1991 Advances in Design Automation*, volume DE-Vol. 32-2, pages 515-522, 1991.
- [12] T. C. Lin and K. H. Yae. Recursive dynamic formulation of a manipulator driven by harmonic drives. *Mechanics of Structures and Machines*, 1993. to appear.
- [13] T. Yoshikawa. Analysis and control of robot manipulators with redundancy. *Robotics Research: The First International Symposium*. The MIT Press. 1984.
- [14] R. Featherstone. *Robot Dynamics Algorithms*. Kluwer Academic Publishers, 1987.
- [16] J. Wittenburg. *Dynamics of Systems of Rigid Bodies*. B. G. Teubner Stuttgart, 1977.
- [15] E. J. Haug and M. K. McCullough. A variational-vector calculus approach to machine dynamics. *ASME J. Mechanisms, Transmissions, and Automation in Design*, 108:25-30, 1986.
- [17] H. Asada and J.-J. E. Slotine. *Robot Analysis and Control*. John Wiley Sons, Inc., 1986.

- [18] C. R. Rao and S. K. Mitra. Generalized Inverse of Matrices and its Applications. John Wiley Sons, 1971.
- [19] C. A. Klein and C. H. Huang. Review of pseudoinverse control for use with kinematically redundant manipulators. IEEE Trans. Systems, Man, and Cybernetics, SMC-13(3):245-250, 1983.
- [20] B. Hannaford and R. Anderson. Experimental and simulation studies of hard contact in force reflecting teleoperation. In IEEE Conference on Robotics and Automation, 1988.
- [21] M. W. Dubetz, J. G. Kuhl, and E. J. Haug. A network implementation of real-time dynamics simulation with interactive animated graphics. In Proceedings of the 12th ASME Design Automation Conference, pages 509-518, 1988.
- [22] J. L. Chang and S. S. Kim. A low-cost real-time man-in-the-loop simulation for multibody systems. In Proceedings of the 12th ASME Design Automation Conference, pages 95-99, 1989.
- [23] J. L. Chang, T. C. Lin, and K. H. Yae. Man-in-the-control-loop simulation for manipulators. In Proceedings of The 3rd Annual Conference on Aerospace Computational Control, pages 688-699, 1989.
- [24] H. N. Christiansen. MOVIE.BYU Training Text, 1987.

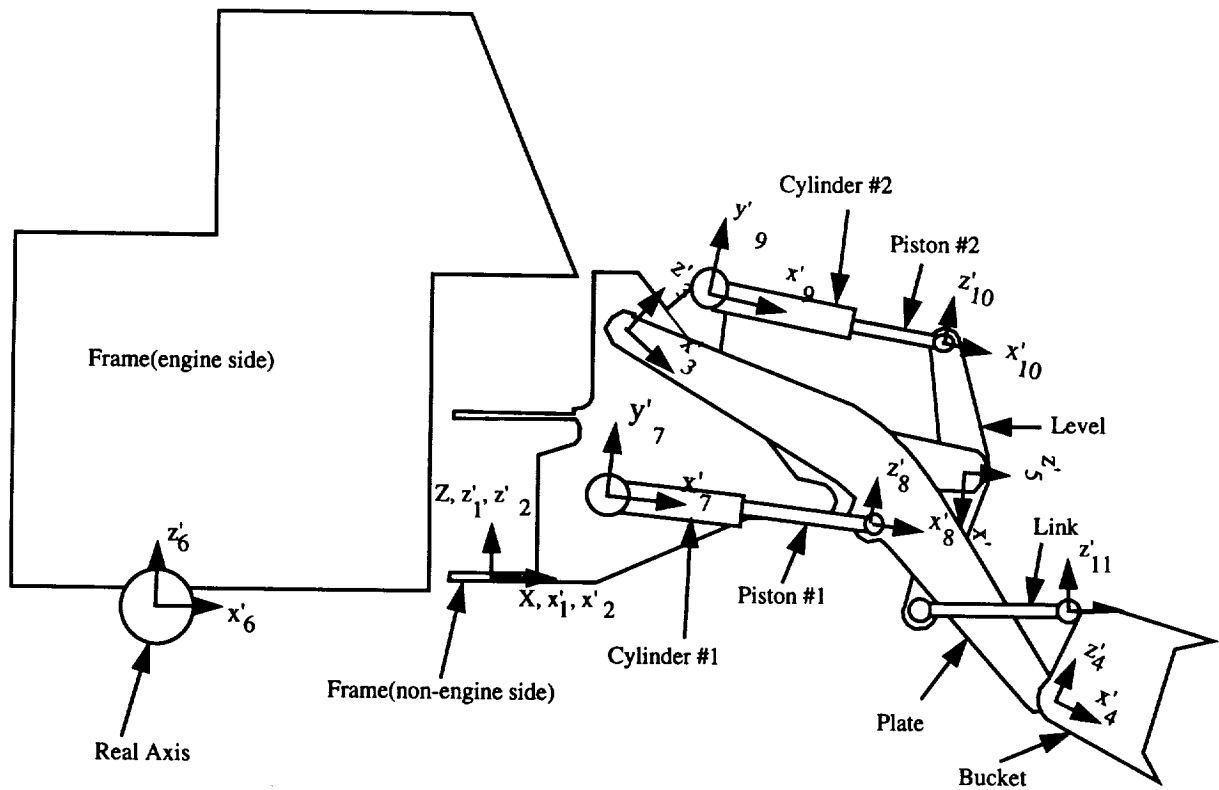


Figure 1: Linkages of Wheel Loader

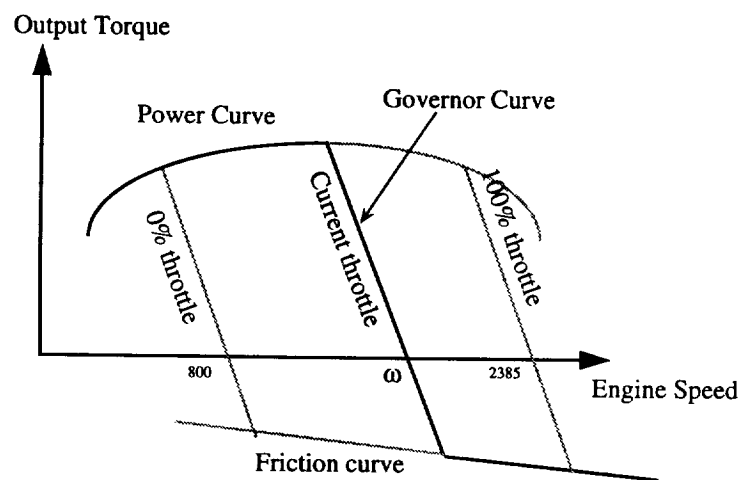


Figure 2: Engine torque output

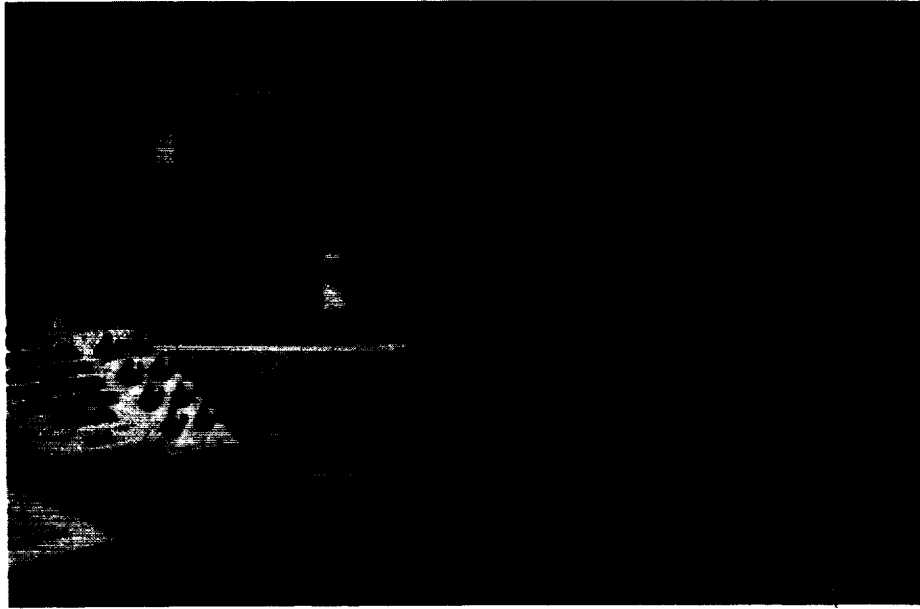


Figure 3: A Kraft mini-master in the foreground and the simulated manipulator in the back-ground on the graphics screen

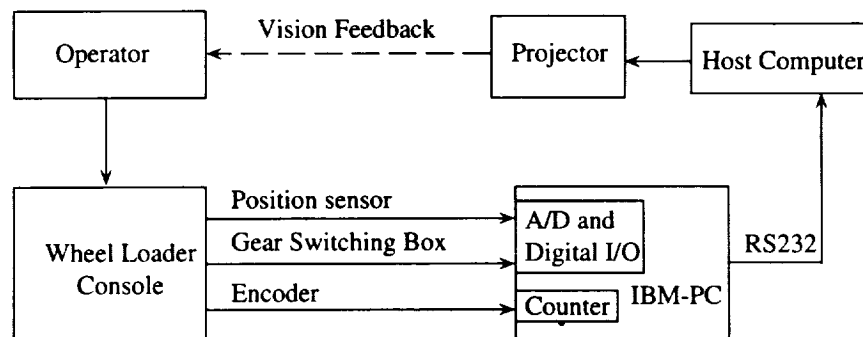


Figure 4: Simulation setup

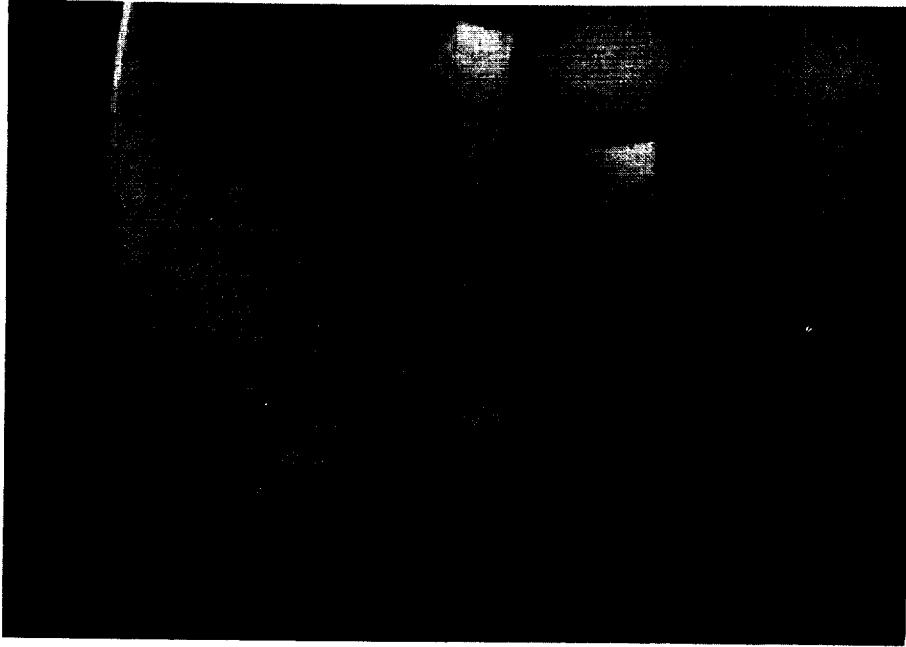


Figure 5: The operator's console and the projection screen



Figure 6: Rotational potentiometers, optical encoders, and their interface wiring

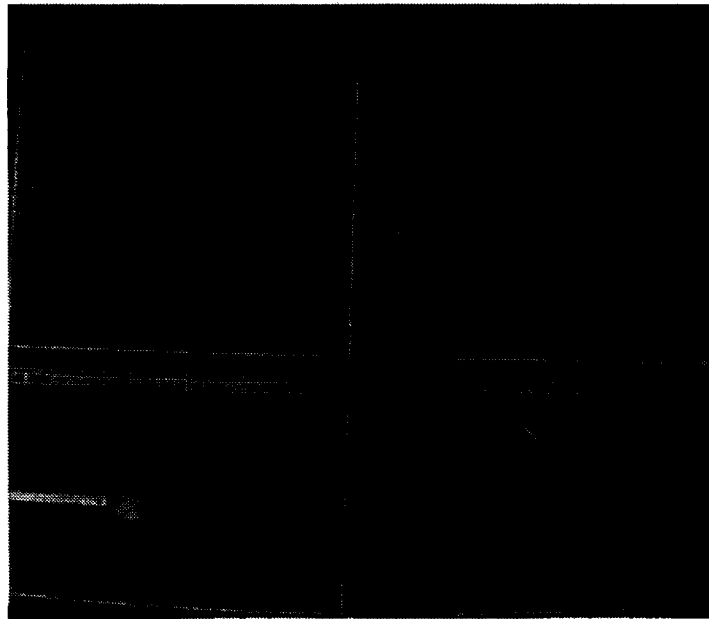


Figure 7: Emulation of camera views on an IRIS graphics workstation: three on the RMS and one on the space shuttle

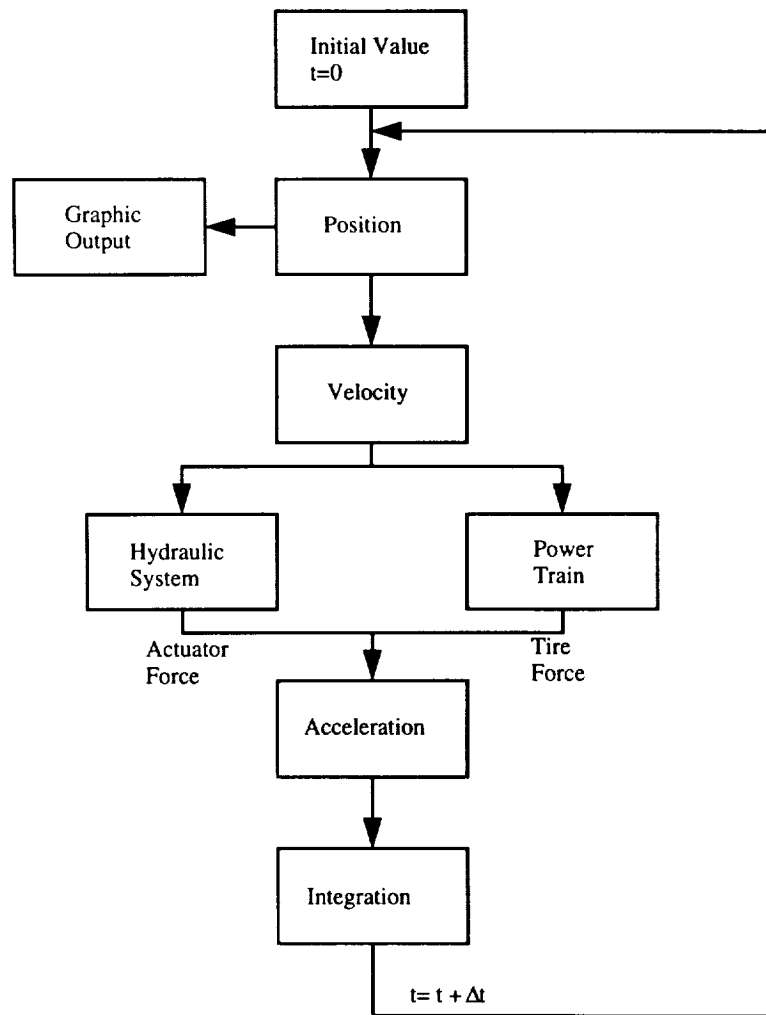


Figure 8: Computation flow

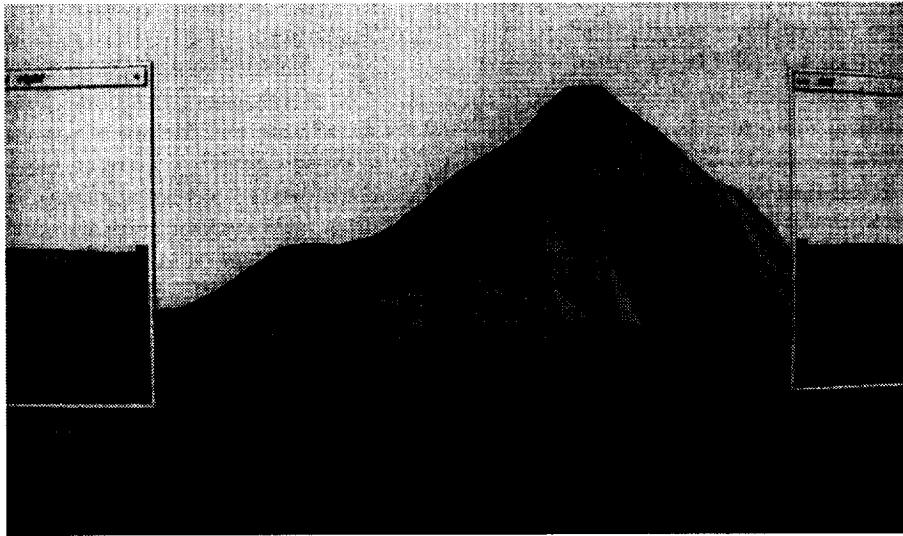


Figure 9: The operator's view displayed with two side mirrors' view

553-09
2536 HIGH PERFORMANCE REAL-TIME FLIGHT SIMULATION AT NASA LANGLEY
P-1

Jeff I. Cleveland II
Project Engineer

National Aeronautics and Space Administration
Langley Research Center
Hampton, Virginia 23681-0001

ABSTRACT

In order to meet the stringent time-critical requirements for real-time man-in-the-loop flight simulation, computer processing operations must be deterministic and be completed in as short a time as possible. This includes simulation mathematical model computation and data input/output to the simulators. In 1986, in response to increased demands for flight simulation performance, personnel at NASA's Langley Research Center (LaRC), working with the contractor, developed extensions to a standard input/output system to provide for high bandwidth, low latency data acquisition and distribution. The Computer Automated Measurement and Control technology (IEEE standard 595) was extended to meet the performance requirements for real-time simulation. This technology extension increased the effective bandwidth by a factor of ten and increased the performance of modules necessary for simulator communication. This technology is being used by more than 80 leading technological developers in the United States, Canada, and Europe. Included among the commercial applications of this technology are nuclear process control, power grid analysis, process monitoring, real-time simulation, and radar data acquisition. Personnel at LaRC have completed the development of the use of supercomputers for simulation mathematical model computation to support real-time flight simulation. This includes the development of a real-time operating system and the development of specialized software and hardware for the CAMAC simulator network. This work, coupled with the use of an open systems software architecture, has advanced the state-of-the-art in real-time flight simulation. This paper describes the data acquisition technology innovation and experience with recent developments in this technology.

INTRODUCTION

NASA's Langley Research Center (LaRC) has used real-time flight simulation to support aerodynamic, space, and hardware research for over forty years. In the mid-1960s LaRC pioneered the first practical, real-time, digital, flight simulation system with Control Data Corporation (CDC) 6600 computers. In 1976, the 6600 computers were replaced with CDC CYBER 175 computers. In 1987, the analog-based simulation input/output system was replaced with a high performance, fiber-optic-based, digital network. The installation of supercomputers for simulation model computation was completed in 1992.

The digital data distribution and signal conversion system, referred to as the Advanced Real-Time Simulation System (ARTSS) is a state-of-the-art, high-speed, fiber-optic-based, ring network system. This system, using the Computer Automated Measurement and Control (CAMAC) technology, replaced two twenty year old analog-based systems. The ARTSS is described in detail in references [1] through [6].

An unpublished survey of flight simulation users at LaRC conducted in 1987 projected that computing power requirements would increase by a factor of eight in the coming decade. Although general growth was indicated, the pacing discipline was the design testing of high performance fighter aircraft. Factors influencing growth included: 1) active control of increased flexibility, 2) less static stability requiring more complex automatic attitude control and augmentation, 3) more complex avionics, 4) more sophisticated weapons systems, and 5) multiple aircraft interaction, the so called "n on m" problems.

Finding no alternatives to using large-scale general-purpose digital computers, LaRC issued a Request for Proposals in May, 1989 and subsequently awarded a contract to Convex Computer Corporation in December of that year. As a result of this action, two Convex supercomputers are used to support flight simulation. The

resulting computational facility provided by this contract is the Flight Simulation Computing System (FSCS). This system is described in references [8] through [11].

ADVANCED REAL-TIME SIMULATION SYSTEM

Through design efforts by both LaRC design engineers and design engineers at KineticSystems Corporation, Lockport, Illinois, three components of the ARTSS were developed to meet LaRC requirements. These were the serial highway network, the network configuration switch, and the signal conversion equipment.

Serial Highway Network

The LaRC ARTSS employs high-speed digital ring networks called CAMAC highways. At any given time, four totally independent simulations can be accommodated simultaneously. The equations of motion for an aircraft are solved on one of the mainframe computers and the simulation is normally assigned one highway. The purpose of the network is to transfer data between the central computers and simulation sites (control console, cockpit, display generator, etc.). The elements of a CAMAC highway are: the Block Transfer Serial Highway Driver (BTSHD); the fiber-optic U-port adaptor, the Block Transfer Serial Crate Controller (BTSCC); the List Sequencer Module (LSM); and the CAMAC crate. Three features of the networks were developed to meet the LaRC requirement. First, the mainframe computer interface to the BTSHD was developed. Second, the block transfer capability was developed to meet LaRC performance requirements. This capability resides in the BTSHD, BTSCC, and LSM. Third, the fiber optic capability was developed to satisfy a site distance problem. The simulator sites are from 350 to 6,000 feet from the computer center.

Prior to the development of the block transfer capability, a CAMAC message was approximately 19 bytes long which included addressing, 24 bits of data, parity information, and response information. The addition of the block transfer capability allowed for the inclusion of many CAMAC data words in a single message. During block transfers, data reads or writes proceed synchronously at one 24-bit CAMAC data word per microsecond. This is several times faster than the normal single word message rate. Besides the CAMAC standard message, there are two modes of block transfer. In the first, the entire block of data goes to a single module within a crate. It is implemented by the BTSCC repeating the module-select and function bits on the crate dataway for each CAMAC word. In the second block transfer mode, the block, either on read or write, is divided among several modules within a crate. This mode employs the LSM module which is loaded by the mainframe computer at set-up time with up to four lists of module-select and function bits. When this type of block transfer is in progress, the BTSCC acquires module number, function, and subaddress for each sequential CAMAC word in the block from the indicated list in the LSM.

To support the high performance requirements for flight simulation at LaRC and Convex supercomputers, a new generation serial highway driver was developed. This driver provides direct connection to VMEbus and allows data to be streamed onto the network. Previous equipment transmitted 24 bits out of the 32 available on the host computer interface; however, the new hardware transmits the full 32 bits from the host computer. Packing/unpacking operations are no longer required to provide the 24 bits in 32 which results in lower input/output latency and increased computer time available for model computation. The new serial highway driver has met the high performance requirements and provides a higher level of programmability.

Network Configuration Switch

The purpose of the network configuration switch system is to provide complete connectability between the simulation applications on the mainframe computer and the various simulation sites. Upon request, any sensible combination of available sites can be combined into a local CAMAC ring network in support of a single simulation. This network configuration for a given simulation is done during the initialization phase, after a highway has been assigned by the scheduling software. The application job requests sites by resource request statements and if the sites are available, the network configuration switch system will electrically and logically configure the network without disturbing other running simulations. The switch is built for a maximum of 12 highways and 36 sites. Each highway may be connected to a different host computer. During the transition

period, four computers were routinely used simultaneously doing flight simulation. Two of these computers were CDC CYBER 175 computers and two Convex supercomputers. In the final configuration, two Convex supercomputers with a total of six configuration switch ports are used.

Signal Conversion Equipment

Three types of output converter modules and two types of input modules were designed and built to LaRC specifications. The converters are high quality and have been added to the KineticSystems Corporation catalog. The digital-to-analog converters (DAC), analog-to-digital converters (ADC), and digital-to-synchro converters (DSC) are 16-bit devices with 14 bits of accuracy. The data transmitted uses 16 bits although only 14 are meaningful. This implementation allows LaRC to change converter precision without major changes in software or protocol. To decrease transmission time, data words are packed such that three converter words (16 bits each) are contained in two CAMAC words (24 bits each). The discrete input converters contain 48 bits per module and the discrete output modules contain 24 bits per module.

Clocking System

Flight simulation at LaRC is implemented as a sampled data system. The equations of motion are solved on a frame-by-frame basis using a fixed time interval. To provide the frame interval timing signals and a clocking system for synchronization of independent programs, LaRC designed and built the real-time clock system. This system is patented and is described in reference [7]. The clock system is composed of a central unit and multiple CAMAC modules called Site Clock Interface Units (SCIUs) which are connected by means of a separate fiber optic star network. Two distinct time intervals are broadcast by the central unit on a single fiber. The first time interval has a constant 125-microsecond period. The tic count necessary for a real-time frame is set in the SCIU by initialization software. This count is decremented by one for each occurrence of the interval timer. When the count reaches zero, each SCIU issues a signal that indicates beginning of frame. The frame time is determined independently for each simulation but must be a multiple of 125 microseconds. The second clock signal, called the job sync tic, has a longer period called the clock common multiple which is set manually, typically in the 1 to 6 second range. This longer period is used for synchronization. Each frame time must divide evenly into the clock common multiple, ensuring that all simulations will be synchronized on the occurrence of the job sync tic.

HIGH PERFORMANCE COMPUTING SYSTEM

The computers that LaRC has put in place to fulfill the requirements are Convex Computer Corporation C3200 and C3800 series computers. These computers are classified as supercomputers and support both 64- and 32-bit scalar, vector, and parallel processing. The first delivery consisted of a Convex C3230 (3 CPUs expandable to 4) with two CAMAC interfaces. The system was delivered with two peripheral buses (PBUS): one PBUS that is used for input/output to standard peripherals such as tape, disk, and line printer and one PBUS that is used exclusively for real-time input/output to the ARTSS CAMAC network. Each VME Input/Output Processor (VIOP) is a Motorola 68020 microcomputer that provides programmable input/output control. Each VIOP is connected to a standard 9U VMEbus and to the corresponding PBUS. The CAMAC interface consists of a KineticSystems Model 2140 Enhanced Serial Highway Driver for VMEbus. The second delivery consisted of one Convex C3850 (5 CPUs expandable to 8) computer configured similar to the C3230 with 3 PBUSs and two CAMAC interfaces. The computer contains 512 megabytes of main memory and sufficient disk and other peripherals to support flight simulation. The C3230 is installed in a secure room and is used for secure simulation and software development.

There are four critical aspects of a computing system to support real-time simulation. These are: CPU performance, memory capacity, time-critical system response, and deterministic system performance.

The first computer installed (C3230) performs a simulation of an X-29 aircraft in 245 seconds per CPU which is 2.7 times faster than the computers being replaced. With two CPUs available for real-time, this results in over 5 times the CPU performance. The second computer (C3840) performs the X-29 in 117 seconds per CPU which is

5.6 times faster than the computers being replaced. With four CPUs available for real-time, this results in over 18 times the CPU performance. The X-29 benchmark has been used to measure performance of other computer systems. Preliminary results of a recent study are shown in Table 1. The reader is cautioned to note that the benchmark measures only scalar CPU performance. The performance index does not necessarily imply any suitability for any application. These results are planned to be reported in a formal NASA publication.

Memory capacity is more than adequate to meet the requirements. The expanded memory capacity, compared with the old system, has allowed LaRC researchers to greatly increase the complexity of the simulations. The increase in memory capacity, coupled with the increase in CPU performance, has led to much higher fidelity simulations. Memory capacity is high enough to permit its use for real-time data storage and retrieval. If the data requirements of the real-time simulations exceed the memory capacity, a disk spooler program will be developed.

Time-critical system response is a measure of how fast the computing system can respond to real-time events from outside the computing system. Time-critical system response on both the computing systems has been measured at 31 microseconds, which exceeds the LaRC requirement.

Deterministic system performance is a measure of how consistently on a frame-by-frame basis the computing system calculates the simulation model without any loss in synchronization with real-time. To use a computing system for real-time simulation, the system must be able to solve the model in a very nearly fixed amount of time, no matter what the demands on the system are for other computing. The C3850 performs simulation models with less than one percent variation in model computing speed. Modifications to the C3230 software are being done to improve its model computing behavior.

Operating System

Convex Computer Corporation offers two real-time operating systems. The operating system currently in use at LaRC requires one CPU for all non-real-time activity: editing, program compilation, and other UNIX activities. The other CPUs may be dedicated to real-time simulation. At the request of a real-time program, the program is locked down in memory to prevent page faults and the CPU or CPUs are dedicated exclusively to the real-time program.

The second real-time operating system incorporates a specially developed real-time kernel that the entire operating system is built upon. With this version of the real-time operating system, the UNIX operating system portion will be pre-empted by real-time requests and the response to real-time interrupts will be deterministic and very short. This version supports, on a single CPU, all activities of a normal UNIX operating system and also simultaneously supports real-time applications.

ADDITIONAL HARDWARE

Processor Communication

LaRC has developed two processor communication CAMAC modules. The Q-bus Microcomputer Interface Module (QMIM) provides an interface between the CAMAC Dataway and a DEC Q-bus on an in-crate microcomputer. The Universal Minicomputer Interface Module (UMIM) provides an interface between the CAMAC Dataway and a DEC DR11-W interface to a compatible computer system. These modules are constructed with two FIFOs, one for input and one for output. They take full advantage of the Look-At-Me CAMAC capability of providing interrupt capability. Input can be done independently of output.

These modules are widely used at LaRC. They provide interfacing to Evans and Sutherland CT-6 image generators, Silicon Graphics devices, Symbolics computers, a variety of DEC computers, and Terabit Eagle 1000 graphics generators.

The contact person for these modules is Don Bennington of LaRC at (804) 864-7353.

Color Target Projector Controller

McDonnell Douglas Corporation has recently developed a color target projection (CTP) system for displaying one target aircraft or other high resolution inset image in a flight simulator projection dome. The system consists of a five axis servo, optical system, and a controller. Two axes, azimuth and elevation, position the projected image on the flight simulator dome surface. Two other axes, zoom A and zoom B, control the image size and focus. A fifth axis, for a neutral density filter wheel, controls the image brightness. The system also has a discrete position solenoid-actuated optical fader disk which may be placed in or out of the projector's optical path to help blend the projected image into the dome's background imagery.

The CAMAC CTP controller module is a circuit card assembly for controlling one CTP from a CAMAC crate. The module occupies one slot in a CAMAC crate and cables directly to the servo amplifier chassis of a CTP system. The module has the following input/output capabilities:

- CAMAC Dataway interface
- two channels of incremental encoder input for CTP azimuth and elevation position servo feedback
- eight channels of 12-bit DAC output for servo amplifier command voltages and test signals
- one discrete output for controlling the CTP fader disk
- one RS-232 serial input/output port

The simulation system host computer communicates CTP servo commands to the module over the CAMAC serial highway and through the CAMAC Dataway. The module inputs CTP azimuth and elevation servo position feedback information from encoders on the CTP body through cabling to the servo amplifier chassis and outputs analog servo command voltages and fader disk position command to the servo amplifier chassis.

The module performs digital position servo control for the CTP azimuth and elevation axes at an iteration rate of 2 KHz. The velocity loops for these two axes are controlled at the analog level through circuitry in the servo amplifier chassis. The position and velocity servo loops for the two CTP zoom axes and the density wheel axis are controlled at the analog level through circuitry in the servo amplifier chassis. The host computer sends position commands to the CAMAC controller module for these three axes and the module relays them to the servo chassis as DAC analog output voltages with no other control action of its own. The host computer also commands the discrete state of the CTP fader disk and the module relays this command to fader disk solenoid driver circuitry in the servo amplifier chassis.

Five of the module's eight DAC channels are required for position servo axis control. The remaining three DAC channels are used by the module to output test signals to aid in system maintenance.

The RS-232 serial input/output port is programmed to communicate with a VT100 compatible terminal. This port is used to display host computer command and status information, monitor CAMAC Dataway transfers, perform programming functions to change capabilities, and perform system configuration and maintenance functions.

Host computer commands for the CTP are sent to the module as blocks of eight CAMAC data words. Each eight word block contains position command values for the five CTP servo axes and fader disk. The blocks also tell the module the rate at which the host computer will send commands to the module. The module is able to receive and process CAMAC command blocks at a rate of up to 500 Hz.

The module is a microcomputer system built using the Intel 80C186EB microprocessor. The software executed by the module is written primarily in Borland Turbo C++ with some assembly language routines. Software development for the module can be done with an IBM compatible PC hosting Borland C++ compilers and assemblers and using Datalight Corporation's C-thru-ROM development tools. The controller can be customized for other applications through programming of the microprocessor. Reference [12] contains additional information. The contact person is Dave Beaman at (314) 234-0386.

CONCLUSION

NASA Langley Research Center has recently completed the development of a system to simulate in real-time increasingly complex and high performance modern aircraft. Utilizing centralized supercomputers coupled with a proven real-time network technology, scientists and engineers are performing advanced research using flight simulation. Hardware and software developed and concepts used are applicable to a wide range of commercial applications that require time-critical computer processing including process control, power grid analysis, process monitoring, radar data acquisition, and real-time simulation of a wide variety of systems.

REFERENCES

1. Crawford, D. J. and Cleveland, J. I. II, "The New Langley Research Center Advanced Real-Time Simulation (ARTS) System," AIAA Paper 86-2680, October 1986.
2. Crawford, D. J. and Cleveland, J. I. II, "The Langley Advanced Real-Time Simulation (ARTS) System," AIAA Journal of Aircraft, Vol 25, No. 2, February 1988, pp. 170-177.
3. Crawford, D. J., Cleveland, J. I. II, and Staib, R. O., "The Langley Advanced Real-Time Simulation (ARTS) System Status Report," AIAA Paper 88-4595-CP, September 1988.
4. Cleveland, J. I. II, Sudik, S. J., and Crawford, D. J., "High Performance Processors for Real-Time Flight Simulation," AIAA Paper 90-3140-CP, September 1990.
5. Cleary, R. T., "Enhanced CAMAC Serial Highway System," presented at the IEEE Nuclear Science Symposium, San Francisco, California, October 23-25, 1985.
6. ANSI/IEEE Standards 583, 595, and 675, Institute of Electrical and Electronic Engineers, 1976.
7. Bennington, D. R., "Real-Time Simulation Clock," LAR-13615, NASA Tech Briefs, June 1987.
8. Cleveland, J. I. II, Sudik, S. J., and Grove, Randall D., "High Performance Computing System for Flight Simulation at NASA Langley," AIAA Paper 91-2971-CP, August 1991.
9. Cleveland, J. I. II, "Application of Technology Developed for Flight Simulation at NASA Langley," presented at the Technology 2001 Conference, San Francisco, California, December 3-5, 1991.
10. Cleveland, J. I. II, "Use of Convex Supercomputers for Flight Simulation at NASA Langley," presented at the Convex Worldwide User Group Conference, Richardson, Texas, May 17-22, 1992.
11. Cleveland, J. I. II, Sudik, S. J., and Grove, Randall D., "High Performance Flight Simulation at NASA Langley," AIAA Paper 92-4179-CP, August 1992.
12. McDonnell Douglas Corporation, "CAMAC CTP Controller Manual", FS-0587, September, 1993.

Computer	Time Seconds	Performance Index*
Hewlett-Packard 9000-735	49	13.5
Cray Y-MP	68	9.7
DEC 3000 Model 500X	76	8.7
HP 9000-720	90	7.3
DEC 3000 Model 500	101	6.5
Cray-2	109	6.1
Convex C3850	117	5.6
IBM RS/6000 970	120	5.5
IBM RS/6000 560	140	4.7
SGI Onyx	176	3.7
SGI Crimson	185	3.6
Convex C3230	240	2.7
CDC Cyber 175	660	1.0

*Note that Performance Index is CPU speed relative to CDC Cyber 175

Table 1
CPU Performance Using NASA Langley X-29 Simulation Benchmark

**THE EFFECTS OF ABOVE REAL-TIME TRAINING (ARTT)
IN AN F-16 SIMULATOR**

**Dutch Guckenberger
ECC International Corporation/UCF
Orlando, Florida**

**Kay Stanney & Norman E. Lane
University of Central Florida
Orlando, Florida**

ABSTRACT

In this application of ARTT, 24 mission-capable F-16 pilots performed three tasks on a part-task F-16A flight simulator under varying levels of time compression (i.e., 1.0x, 1.5x, 2.0x, and random). All subjects were then tested in a real-time (1.0x) environment. The three tasks under study were an emergency procedure (EP) task, a 1 versus 2 air combat maneuvering task, and a stern conversion or air intercept task. In the EP task, all ARTT pilots performed the EP task with 28% greater accuracy, and were better at dealing with a simultaneous MIG threat, reflected by a six-fold increase in the number of MIG kills compared to a real-time control group. In the stern conversion task, there were no statistical differences between group. In the ACM task, those pilots trained in the mixed time accelerations were faster to acquire lock, and were faster to kill both MIG threats than the other groups.

These findings are generally consistent with previous findings that show positive effects of task variation (including time variations) during training. Also discussed are related research findings that support the benefits of ARTT, and ARTT's impact on emergency procedure training. Further, a synthesis of multi discipline research outlining the underlying theoretical basis for ARTT is presented. A proposed model of ARTT based on an analogy to Einstein's theory of special relativity is suggested. Conclusions and an outline of future research directions are presented. Successful current commercialization efforts are related as well as future efforts.

INTRODUCTION

Above Real-Time Training (ARTT) refers to a training paradigm that places the operator in a simulated environment that functions at faster than normal time. In the case of air combat maneuvering, a successful tactical air intercept which might normally take five minutes, would be compressed into two or three minutes. All operations of the intercept would correspondingly be accelerated such as airspeed, turn and bank velocities, weapons flyout, and performance of the adversary. In the presence of these time constraints, the pilot would be required to perform the same mission tasks to the same performance criteria--as he would in a real time environment. Such a training paradigm represents a departure from the intuitive, but not often supported, feeling that the best practice is determined by the training environment with the highest fidelity. ARTT can be implemented economically on existing simulators. It is important to realize that ARTT applications require the simulated velocity of the targets and other entities to increase, not the update rate. Over 25 years ago, NASA Dryden's flight test engineers recognized that if one could program a simulator to operate in "fast time", one could give test pilots a more accurate experience or "feel" of real-world stresses that would be present in the aircraft [1] [2].

The bulk of support for ARTT, in simulators, comes from anecdotal reports from NASA. Researchers at the NASA Dryden Flight Research Center during the X-15 program in the late 1960's needed a mechanism to address the X-15 test pilots' post flight comments of being "always behind the airplane..." and "... could never catch up" [3]. Clearly, there were some differences between the perceived time in the well-practiced simulator flights and perceived time in the experimental aircraft. NASA Dryden's Jack Kolf originated the fast-time simulation concept and the first time NASA used fast time simulation was toward the end of the X-15 program.

Pilots compared practice runs at various time constants with flights they had already flown. A fast time constant of 1.5x felt closest to their flight experience and was planned on being implemented in the lifting body programs, but lack of funding precluded the program from fully developing the capability. Regardless, NASA's test pilots at DFRC have endorsed the use of "fast time" simulation as part of the training process[1] [2].

Vidulich, Yeh, and Schneider [4] examined the utility of time compression as a training aid for training a basic air traffic control skill (a high performance skill) [16]. One group practiced the intercept with the target plane traveling at 260 knots. The second group practiced the intercept at 5200 knots - 20 times real time! The subjects in this group received between 72-80 trials per hour during training. Both groups were then tested in real time. The time compressed group was significantly better at identifying the turn point; there was no difference between groups on estimating roll out heading for the intercept.

Guckenberger, Uliano, and Lane [5], using a table top tank gunnery simulator, trained naive subjects on three tank gunnery scenarios under five acceleration factors (i.e., 1.0x, 1.5x, 2.0x, sequential, and mixed). Their results demonstrated that training time could be cut up to 50% with performance staying equal to or surpassing a real-time control group. Further, in one ARTT group (mixed presentation) their mean performance scores were 50% higher than the control group (1.0X).

Commercialization of ARTT into the mainstream is already being implemented by nine U.S. companies, see future research directions for details.

THEORETICAL UNDERPINNINGS

Psychophysical research into time perception has shown the relativistic nature of time perception in humans [8] [9] [10]. Relativistic nature is defined as linking a human observers perception of time to that particular observer's "stimulation state" or "time norm" analogous to Einstein's theory of special relativity linking relative velocities to a particular observer frame of reference norm. It is noteworthy that this analogy was arrived at independently by Jones [8], Guckenberger [5] and Toumodge [10] from three different fields. Hahn and Jones have even developed working models [11] though their work is primarily in the area of Audio training. Dr. June Skelly is attempting to extend the Audio finding to the arena of Visual training and has already generated some impressive initial results [8]. Brevity of this paper format precludes further in depth synthesis of multi discipline research to support the theoretical basis for ARTT, suffice it to indicate that ARTT now has a firm theoretical basis upon which to build. The foundations for ARTT and Human perception are well established. Time perception can be altered if a particularly boring or interesting task is introduced, or if the arousal state of the subject is changed through external environmental cues [12]. Humans perceive time differently depending upon the individual's "stimulation state" or "time norm" This stimulation state is based, in part, on the sensory cues in the environment and the interactivity level between the individual and his/her environment. Perceived time, therefore, is tied to the particular individual at his/her particular stimulation state to form a "time frame of reference" for that individual. Cohen [13] discusses evidence for an interrelationship between one's "inner clock" and sensory/motor functioning where each can influence each other to alter the perception of time. Most high performance tasks involve both sensory/motor and cognitive skills. Further Wright-Patterson Researchers have developed a method of Rapid Communication (RAP-COM) which improved throughput and retention [14].

When this subjective time reference is perceived as long, it may offer a unique advantage for providing training on critical high performance skills. This artificially accelerated frame of reference may give the operator more "time" in which to actually perform key elements of the mission. It is important to note that when using ARTT more compressed training trials can be performed in the same amount of time. The very realization that the operator has more time may lead to better decision making and situational awareness. It may give the operator the edge that makes the difference in today's modern battlefield. More training trials per unit time is reason enough to implement ARTT. As long as no negative training is introduced, more economic training can occur on existing simulators. The simplest case for ARTT is improved simulator usage either by more trials per unit time per trainee, or higher trainee throughput. Recent experiments extending ARTT to virtual reality have shown ARTT produces higher performance, reduced frustration and stress, reduced temporal workload using validated NASA Wewerinke TLX scales [7].

RESEARCH OBJECTIVES AND HYPOTHESES

The objectives of this task is to conduct research regarding: (1) the relative effectiveness of ARTT versus conventional training on different simulator platforms; (2) the relative effectiveness of alternative implementations of ARTT; and (3) the impact of ARTT versus conventional training on total time. Prior research suggests that training in a time accelerated environment should lead to poor performance versus a control group, but should lead to greater performance on a real-time transfer task. Second, it is expected that there will be group differences in training as a function of the time acceleration constant that is used. Third, it is obvious that training time will be reduced in direct proportion to the time acceleration constant used. Finally, it is not expected that training under various time manipulations will lead to negative transfer of training to a real-time task.

METHOD

Subjects

Twenty-four mission-capable F-16 Air Force pilots from the 56th Tactical Training Wing, MacDill Air Force Base, Tampa served as subjects for this experiment. This subject pool had 743 mean flight hours (range of 300-3400), and 134 mean simulator hours (range of 30-500). All subjects were recruited on a voluntary basis in accordance with American Psychological Association (APA) Principles for Research with Human Subjects. Prior to testing, subjects were given written instructions informing them as to the general nature of the experiment.

Equipment and Materials

Two Avionics Situational Awareness Trainers (ASAT) were used as the testbed for this study. The ASAT is a low-cost F-16A cockpit trainer designed primarily to train in the beyond visual range (BVR) environment. The hardware components that make up the ASAT consist of three personal computers (PCs). The host computer is a PC-AT with an i386 CPU and a i387-20 co-processor, which drive the head-up (out-the-window) and radar electro-optic (REO) displays and collect the data coming from the stick and throttle. Another PC-AT computer (i286), drives the radar warning receiver display. Sound and vibrational cues are provided through the third PC which drives a stereo amplifier, seat and back cushion-mounted speakers, and sub woofers. Aural cues available in the ASAT include radar sensor tones, engine and air noise, missile launch, and gunfire, radar warning receiver (RWR) tones, and missile seeker head tones.

Graphics for the head-up display are high resolution, 1024 x 1024 RGB, with a 63.36 kHz horizontal scanning frequency. The monitor for the head-up and visual display is a 19-inch color CRT monitor which is mounted in front of the pilot on top of the cockpit enclosure, and gives the pilot a 23° X 23° field-of-view. The REO display simulates that of the F-16A Block 15S AN/APG 66 radar, and is presented on a 5" monochrome monitor. It is driven by the i386 and is controlled through switch activation on the throttle and by a radar control panel located on the left side of the simulator. The panel contains active switches to control antenna azimuth, antenna elevation and target history selection. The radar warning receiver (RWR) simulates the ALR-69 RWR, and the display consists of a 9" EGA resolution color monitor. All symbology is generic and unclassified.

The side-stick controller and throttle are high fidelity copies of the controls used in the actual F-16A. The stick can experience a maximum deflection of 0.25" in each of the four axis (forward, backward, right, left), and is equipped with buttons that allow the performance of different functions which include four way trim, missile release, gun triggering, missile select button (AIM 9-J/L), and a return to search switch. The throttle controls thrust from idle to full military power and beyond through five stages of afterburner. (It should be noted that no change in thrust results in the ASAT from afterburner stage 2 through stage 5; the afterburner has only two states: on and off.) Other throttle functions include: four way radar cursor, UHF/VHF transmit switch, missile uncage button, speed brake switch, antenna elevation knob, chaff/flare release button, and dog fight switch.

The ASATs communicate via a PC-based ethernet network at the asynchronous rate of approximately 10-14 packets per second. For the purpose of this experiment, the network was modified so that each ASAT communicated through a Hewlett-Packard i386, 33 MHz PC which served as the experimental interface.

This PC controlled task selection, trial start and stop times, duration, data storage, and other experimental information. In this design, the PC would also send messages to either ASAT instructing the simulator to activate or deactivate certain functions (e.g., sound) that were required for a subject to perform a given task. Special purpose C and assembly software was written to handle these special requirements

Procedure

The subjects' first mission was to familiarize themselves with the simulator, including its displays, controls, and handling qualities. These aspects of the simulator are probably different than what the subjects are normally accustomed to. Since the F-16A model is no longer in service with the U.S. Air Force, only some of our subjects had ever flown it. Based on preliminary test subjects, we do not believe this to be a problem since the F-16A and F-16C models have sufficiently similar aerodynamic and avionics characteristics. The subjects were given approximately forty-five minutes for familiarization across a wide variety of scenarios. During this time, the subjects were encouraged to experiment with the controls, displays, and the flying characteristics of the simulator.

After the familiarization period there was about a fifteen minute break. The subjects then flew an assigned order of the three tasks at an assigned ARTT value. These assignments have been made beforehand and represent a complete counterbalancing of the four ARTT conditions, three tasks, and 24 subjects. For each task, the subject flew 10 trials at the assigned group, subjects were presented with a random presentation of the first three time constants. The within-group factor tested a trial effect with each subject receiving 10 training and 4 test or transfer trials. Dependent variables included varied flight performance data such as time-to-lock, time-to-kill, hit/miss percentage, mission performance times, and emergency procedure checklist performance. Specific data collected were a function of the task being performed

Training Tasks and Initial Conditions. The three tasks used for this study are listed and explained below. A task ended when the subject "killed" the target(s) or when the task timed-out. We limited any given task to five minutes to optimize data storage. For each hop for each task, the subject had unlimited fuel. The subject did not have access to any ground control intercept (GCI) or airborne AWACS information. The following task briefings were the only information available.

Task 1 - One versus Two Air Combat Maneuvering. Two bogeys on the nose at 25,000 ft. Goal was two valid face shots on the initial merge. Continue to engage the bogeys until they have been killed, or until the experimenter terminates the hop.

Task 2 - Stern Conversion. Bogey was 40 miles on the nose at 20,000 ft. Goal was to perform stern conversion and position for a possible AIM 9J missile or gun shot as quickly as possible. Maximum distance for weapons employment was 1500 ft. The subject was required to maintain a 30 degree aspect cone at no more than 1500 feet before permission to fire was given. This allows for adequate data collection. This hop ended when the bogey has been killed or when the experimenter terminates the hop.

Task 3 - Emergency Procedure. In this task, the subject was flying over enemy area suspected of having energy pulse weapons (better known as "power sucker"). The subject must deal with two external threats. Namely, the "power sucker" and an enemy bogey. When the subject was painted by one of these weapons, he heard (and felt) a constant low rumbling noise indicating an imminent and catastrophic power loss. If this happened, the emergency procedure (EP) to defeat this weapon was as follows:

- 1) fire energy decoy (missile);
- 2) change heading left 10 degrees; 3) hit energizer (flare);
- 4) change heading right 10 degrees; 5) fire energy decoy (missile);
- 6) hit energizer (flare).

If the subject performed the procedure above exactly, and in the correct order, the "power sucker" would be defeated and aircraft power would be restored. If not, the subject would crash. The goal of this task was to perform the EP above as quickly as possible while at the same time successfully engaging a hostile bogey.

RESULTS

Raw flight performance data originally collected at a 10-14 Hz iteration rate were reduced into trial summaries. Summary data were then analyzed using the Statistical Package for the Social Sciences (SPSS) [15]. The multivariate analysis of variance (MANOVA) syntax for SPSS was used as the overall design structure for the analysis; however, univariate F tests were calculated for specific planned comparisons of interest. These planned comparisons focused on identifying statistically-reliable differences between the performance of the four time acceleration groups in training, and performance comparing the average of the three training blocks (for a given task/dependent variable combinations) with the two transfer trial blocks.

For the emergency procedure (EP) task, number of MiGs killed, time to complete EP, and percent of EP performed correctly were analyzed by group. Analysis of the EP flight data demonstrated a significant increase in MIG kills from training to transfer for all accelerated conditions ($F_{3,20} = 10.87, p < .01$) with the 1.5x and 2.0x conditions slightly outperforming the mixed group. The three accelerated groups, at the conclusion of the last transfer block, had a better than six-fold advantage in the number of MIG kills compared to those trained at real-time (see Figure 2 on next page.) Further, the EP results demonstrated that all the groups trained under accelerated time conditions produced significantly higher accuracy in performing an emergency procedure in the transfer condition than did a real-time control group. The mixed and the 2.0x groups performed the EP near perfectly (100% and 96.6%, respectively). The 1.5x group's accuracy was almost 90%, while the control group scored the lowest at about 72%. (see Figure 1 on next page.)

When comparing performance in training on the number of MIG kills, there is also a significant difference between the groups ($F_{3,20} = 3.95, p < .05$). Both the 1.5x and 2.0x groups performed better in training when compared to the 1.0x and mixed groups. This finding was not expected, and is not consistent with what is known about the contextual interference phenomenon.

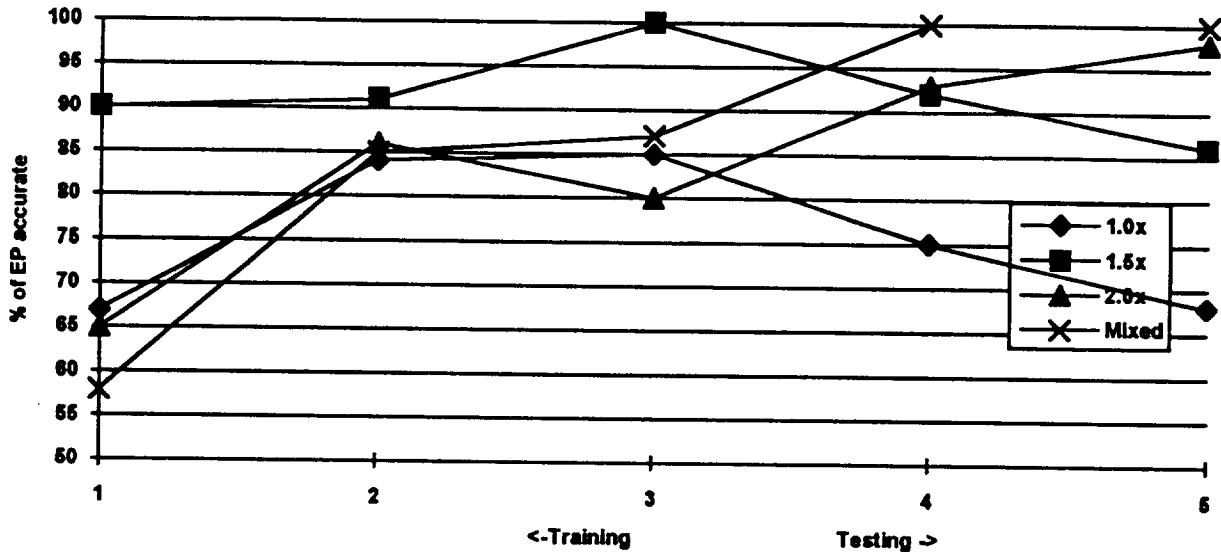


Figure 1. Mean percentage of EP performed correctly by trial block
Block 1..3 = training, Block 4..5 testing

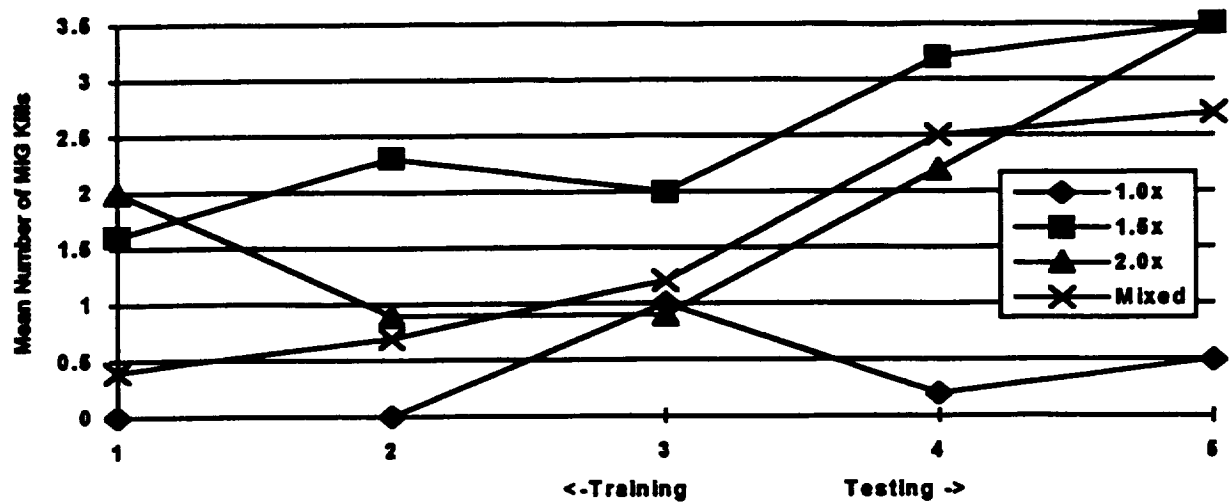


Figure 2. Mean Number of MiG Kills by trial block
Block 1..3 = training, Block 4..5 testing

Next, the time to complete the EP procedure, and percent of EP procedure performed correct were analyzed. As time went on, all the groups completed the EP checklist items quicker, although that difference was not statistically reliable. When comparing the accuracy performance, however, both the 2.0x and mixed conditions performed the checklist task significantly better than either the 1.0x or 1.5x groups, when later tested at real-time ($F_{3,20} = 7.45$, $p < .002$). In fact, subjects in the mixed group scored perfectly in the transfer condition. The 1.0x and 1.5x groups actually saw a slight decrease in accuracy performance from training to transfer. There were no mentionable differences between the groups in training.

For the stern conversion task, time to reach criterion, stern score, and distance at lock were analyzed by group. Analysis of the stern conversion task showed that the 1.5x group performed only slightly better than the other groups in the time to reach a preset position criterion. The 1.5x group performed the task faster in training *and* in transfer but the reader will note that these findings are not statistically significant.

For the distance at lock variable, which represents a measure of radar target acquisition performance, the 2.0x and 1.5x groups performed slightly worse in training, indicating that subjects in those two groups took somewhat longer to locate and lock the bogey. With this variable, the greater the range at which the bogey is identified and locked, the better opportunity a pilot has to make decisions. In transfer, the 1.0x and 1.5x groups continued to improve, however, the mixed group showed a significant decrease in the first transfer trial block ($F_{3,20} = 37.64$, $p < .001$) (see Figure 3). This latter finding could be due to the relative uncertainty of the initial closure speeds and range-to-target caused by mixing the accelerated conditions.

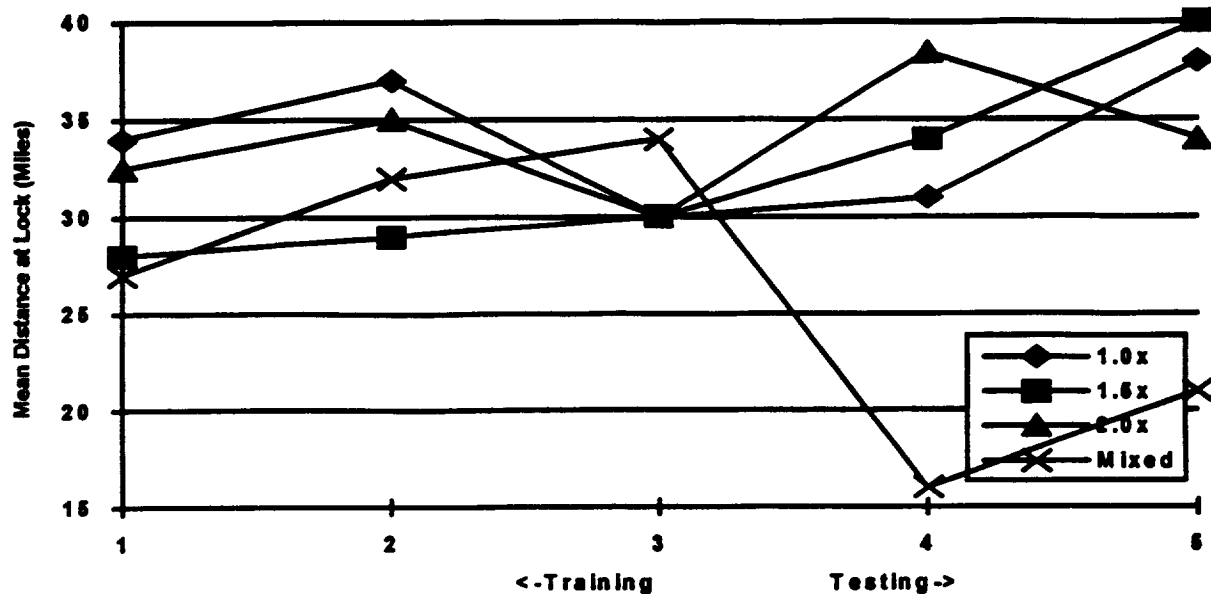


Figure 3. Mean distance at lock by trial block
Block 1..3 = training, Block 4..5 testing

For the stern conversion score, there are no significant differences between groups in training or between training and transfer performance among the four groups. The scoring procedure used for the stern task is based on a subjective rating that is often given by instructor pilots (IPs) to students. The score is based on assessing both the closure speed and aspect angle during the conversion. The idea being that when the pilot rolls-out behind the bogey (low aspect angle), the pilot should not be more than three miles or less than one mile behind the bogey. As a rule-of-thumb, the closure speed should also be in proportion to the distance (e.g., at 2 miles, 200 knots closure speed). Although not statistically different, there is an actual decrease in performance from the last training block to the first transfer block followed by a slight increase in performance at the last transfer block. In the end, performance for the 1.0x group is higher than the other groups. The results of the stern conversion, taken together, tends to suggest that piloting tasks that involve well-learned (at real-time) and continuous responses to both internal (ownship) and external (bogey) positioning cues might not benefit from above-real-time simulation.

For the air combat maneuvering (ACM) task, time to first lock, time to reach criterion, and number of valid missile shots were analyzed by group. For time to first lock, which is a measure of the speed at which a pilot acquires his adversary on radar, all groups except the 1.0x group saw a significant increase in lock time from the last training block to the first transfer block ($F_{3,20} = 2.92, p < .05$). In comparing the groups at the final transfer block, both the mixed and 1.0x groups performed significantly better than either the 1.5x or 2.0x group. The 2.0x group also outperformed the 1.5x group in transfer (see Figure 4).

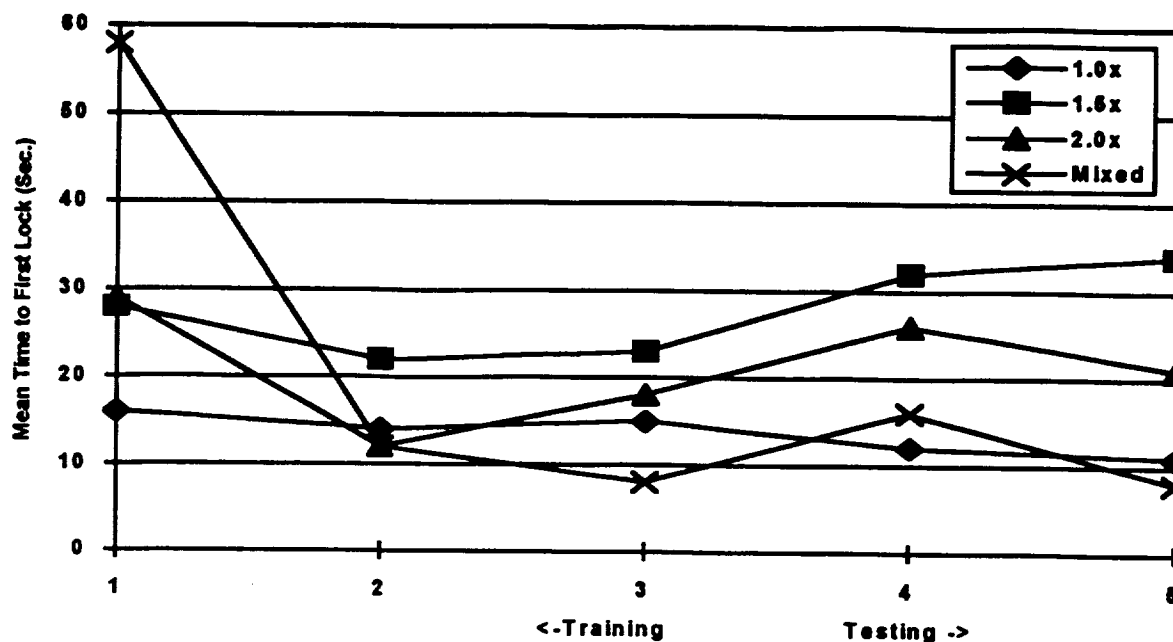


Figure 4. Mean time to first lock by trial block
Block 1..3 = training, Block 4..5 testing

For the time to reach criterion, there was no significant difference between groups from training to transfer. In comparing the last transfer block, however, the mixed group performed significantly better than either of the other groups ($F_{3,20} = 4.55, p < .014$) (See Figure 5).

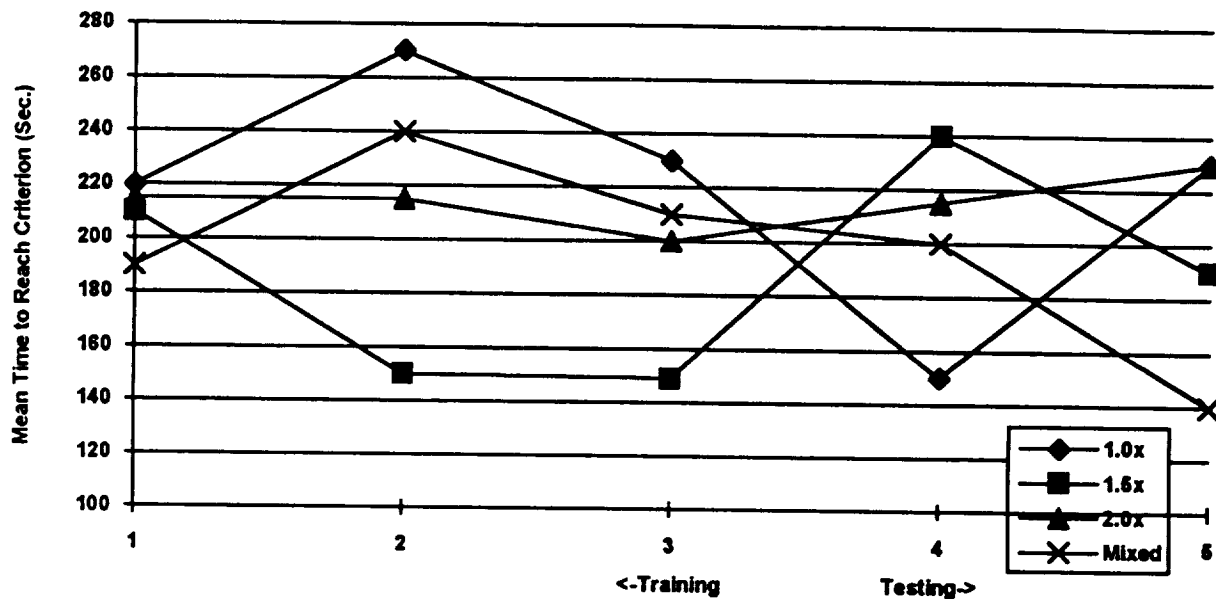


Figure 5. Mean time to reach criterion by trial block (ACM)
Block 1..3 = training, Block 4..5 testing

Finally, the mean hit/miss percentage were analyzed and revealed no significant differences between group in either training or transfer. Upon further inspection, it was apparent that this metric was somewhat biased due to the performance of the missiles. This point is expanded in the discussion section below.

DISCUSSION

The EP results demonstrated that all the groups trained under accelerated time conditions produced significantly higher accuracy in performing an emergency procedure in the transfer condition than did a real-time control group. The mixed and the 2.0x groups performed the EP near perfectly (100% and 96.6%, respectively). The 1.5x group's accuracy was almost 90%, while the control group scored the lowest at about 72%. This finding in particular demonstrates that ARTT may have potential to train procedural tasks with greater accuracy and in less time. In the EP task, the difficulty of the task was increased by placing all groups under the additional (simulated) stress of having to perform the EP during a secondary air combat task. An unexpected result was each ARTT group, the number of enemy MiGs killed was six times higher than the 1.0x groups when compared in the real-time transfer blocks. There was also no significant difference between the groups when analyzing the time to complete the EP variable. The subjects, after a few trials, mastered the procedure and their performance stabilized. This seems to indicate that ARTT does not necessarily effect the speed with which pure motor tasks are performed, rather ARTT benefits the internal decision making process.

Results of the stern conversion tasks are less clear, and neither support or refute the ARTT concept. For this task we attempted to implement ARTT by increasing the velocities of the ASAT and the bogey. In retrospect, due to the physics and geometry of the stern task, we failed to create a savings or reduction in training time which is a central tenet in ARTT. The task forced the ARTT groups to take essentially the same time in training as the real-time control group. In other experiments we have been successful by speeding up targets, ownship, or both. This was not the case for the stern task. Moreover, pilots differ greatly in their approach to performing the task. Some would perform a low/high or high/low vertical conversion while some would initially offset left or right and perform a "standard" conversion. This made it difficult to establish useful measures of performance. Tasks such as the stern conversion that could be performed successfully using one or more alternate strategies, did not produce useful measures.

The air combat maneuvering (ACM) task also produced mixed results. Again, the fact that pilots have different flying styles leads to difficult performance assessment. The pilots were instructed to take two valid face shots - one at each bogey. A "valid" shot was one in which the range from the bogey was less than or equal to six miles and the aspect angle was between 135 and 180 degrees. The ASAT software modeled only the older AIM-9J and AIM-9L missiles. Unfortunately, when the raw data were inspected, it became clear that the pilots had great difficulty achieving "valid" missile shots, as they were defined, regardless of the group they were assigned. The explanation for this phenomenon lies in the performance of the missiles and the attack profiles preferred by the pilots. Specifically, the AIM-9L is capable of a high aspect kills, but its performance is significantly worse than the newer AIM-9M which the pilots are familiar with. The hit/miss percentage metric, therefore, cannot be considered a true reflection of pilot/weapon performance. In addition, most pilots chose to "offset" or break right or left to create more of an advantageous aspect angle. With a less than optimal high aspect kill performance of the AIM-9L missiles, the fight usually degenerated into a tail chase with a time savings disappearing since both the ASAT and the MiGs were both accelerated.

There were some trends in the ACM task that, although are not statistically significant, bear some mentioning. The mixed group were 11% faster in disposing of the two MiGs. The mixed group also showed the fastest reduction in time to first lock from training to transfer. Finally, the hit/miss percentage score was highest in the 1.5x and 2.0x groups.

CONCLUSION

Based on the results of this research, tasks that contain simple psychomotor or procedural components such as the emergency procedure task performed on the F-16 ASAT clearly benefit from ARTT. Moreover, this research demonstrated that task type and task content are differentially affected by ARTT. The ARTT groups showed higher performance scores when compared to a real-time control group in transfer for the EP task. For tasks with more complex cognitive components such as the ACM and stern conversion, there was no clear advantage in the ARTT groups compared to a real-time control group. The stern and ACM tasks allowed for alternative performance strategies that pose particular measurement and interpretation problems.

ARTT potential benefits for emergency procedure training can not be over emphasized. The increase accuracy of performing EPs bears further study because of the obvious implications for safety training. Many real-world emergencies require accurate performance of checklist procedures under sometimes extremely stressful circumstances. In this study those trained under an accelerated condition not only performed the primary EP more accurately, they also were able achieve a significantly greater number of MIG kills (6x) on a concurrent secondary task.

ARTT obviously has utility in the weapons training process, further, consider it may be possible to accelerate a pilots "time norm" in the cockpit just prior to combat the ARTT pilot advantage in the time dimension will increase combat effectiveness and situational awareness.

With respect to the initial research objectives:

1. ARTT was more effective than conventional real-time training in the case of EP task. The stern conversion and ACM task results were mixed.
2. For those significant effects, the group that provided the greatest performance improvements was the one that mixed the presentation at different speeds. This supports the contention that task variety in training leads to higher performance.
3. The impact of the ASAT study on training time is inconclusive due to methodological considerations.

Finally, as expected none of the ARTT groups experienced any negative transfer of training to real-time transfer tasks.

The results of this experiment can be seen as further support of the benefits of training at Above-Real Time. The emergency procedure task results illustrate the performance increases obtainable using ARTT on existing simulators. The other two tasks did not restrict the pilot's actions sufficiently to allow useful measures to be obtained. American pilots are arguably the finest pilots in the world, but their independence and cunning that make them great, also makes them difficult to restrict and measure. Consider the evolution of research listed below:

- The first use of "fast time" or ARTT in simulators was Jack Kolf at NASA Dryden over 20 years ago. [1]
- NASA'S initial success was followed by successes in the lifting body program as well. [2].
- The success of ATC by the FAA study. [4]
- Success of VIGS time saving and performance increase. [5]
- Emergency procedure in F-16 accuracy increase [6]
- Virtual time in VR reduced stress and workload [7]

Applications of ARTT to simulators seems to have merit. The theoretical frame work for ARTT continues with synthesis from many diverse fields, most notably audio perception who's relativistic working models may transfer to illuminate ARTT's working relativistic model.

ARTT and the intrinsic time adaptability of man is a vast field of great potential.

FUTURE RESEARCH DIRECTIONS

Near-term work will focus on expanding the application of ARTT for emergency procedure training. We are also beginning to explore techniques to test the effectiveness of ARTT on subsequent performance in the actual aircraft.

The overall aim of the ARTT concept is to exploit the time adaptability of humans and foster a new way of thinking about time manipulation in the man-machine interface. Future research directions might include safety, education, medical, and entertainment applications. For example, it would be possible to increase the voice and data communication rate over a network to allow crews or teams to train at faster than real-time. Also, as scientists explore the concept of ARTT and virtual time, the real world bond we have with perceived time

will weaken. Time flow could be controlled for the benefit of the trainee. New training methods that are time flexible will change form, fit and function of the man-machine interface. ARTT programs are initially planned in simulation and training with follow on to use of ARTT in other man-machine interfaces. Emergency procedure training for pilots, both commercial and military is envisioned as the initial proving ground.

Current and near future Research Projects include:

- ARTT for airborne weapons training
- Virtual Time Adding the fourth Dimension to Virtual Reality: Next Generation Man-Machine Interfaces
- Above Real Time Communication
- ARTT applications in a DIS environment
- ARTT applications in Video Decompression
- ARTT Theoretical model: Relativistic Time-Speed Reading-Speed Listening -> Speed Simulating
- Time adaptive training and time adaptive human computer interfaces
- Slower than real-time in the human-computer interface to benefit the elderly, disabled and disadvantaged

Commercialization of ARTT into the mainstream is already being implemented by nine U.S. companies, see below for outline of details.

ARTT commercialization has begun:

- ECC has modified six different simulators to include the technique
- Silicon Graphics Flight, Shadow and Dogfight simulations now support ARTT
- Coryphaeus Designers' Workbench now supports the ARTT interface
- Pellucid's OPEN-GL library is planning support
- Link, fellowship for advanced simulation and training has been awarded to further ARTT research
- TWA is considering participating in a pilot program
- Total Quality Tennis has contracted to build ARTT simulators
- The Chicago Cubs are negotiating to improve batting through ARTT simulation
- Most importantly, NASA and the Airforce are seeking to support further research for the use of ARTT in emergency procedures training and improving air safety

REFERENCES

- [1] Kolf, J. (1973). *Documentation of a simulator study of an altered time base*. Unpublished manuscript.
- [2] Hoey, R. G., (1976). *Time compression as a means for improving the value of training simulators*. Unpublished manuscript.
- [3] Thompson, M. (1965). *General review of piloting problems encountered during simulation flights of the X-15*. Paper presented at the 9th Annual Meeting of the Society of Experimental Tests Pilots.
- [4] Vidulich, M., Yeh, Y.Y., & Schneider, W. (1983). *Time compressed components for air intercept control skills*. Proceedings of the 27th meeting of the Human Factors Society. 161-164.
- [5] Guckenberger, D., Uliano, K.C., & Lane, N.E (1992). *The Application of Above Real-Time Training for Simulators: Acquiring high performance skills*. Paper presented at the 14th Interservice/Industry Training Systems and Education Conference, San Antonio, TX, , 928-935
- [6] Guckenberger, D., Uliano, K.C., & Lane, N. E. (1993). *Training High Performance Skills Using Above Real Time Training*. NASA Final Report Nag # 2-750
- [7] Guckenberger, D., & Stanney, K. (1993). *Virtual Time: Adding the Fourth Dimension to Virtual Reality*. Paper presented at the 15th Interservice/Industry Training Systems and Education Conference, Orlando, Fl.

- [8] Jones, Mari R. Time Our Lost Dimension *Psychology Review* 1976 Vol. 83 (5) 323-355
- [9] Skelly, June J. (1993) *The Role of Event Time in Attending Time & Society* Vol. 2(1) , 107-128
- [10] Toumodge, S.S. (1990) *Signal and Data Processing of Small Targets* Paper presented at the Orlando April 1990 Annual Meeting of the Society of Photo-Optical Instrumentation Engineers. (A91 - 36901 15-32) 378-385.
- [11] Hahn J. & Jones M. J. (1981) Invariants in Auditory Frequency Relations, *Scandinavian Journal of Psychology* 22, 129-144.
- [12] Parasuraman, R. (1986). Vigilance, monitoring, and search. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance: Volume II: Cognitive processes and performance* (pp. 43-1 - 43-39). New York: Wiley.
- [13] Cohen, J. (1964, November). Psychological time. *Scientific American*, pp. 116-124.
- [14] Matin, E., & Boff, K.R. (1988). Information transfer rate with serial and simultaneous visual display formats. *Human Factors*, 30, 171-180.
- [15] SPSS, Inc. (1992). *SPSS for Windows* [Computer Program]. Chicago, IL: SPSS, Inc.
- [16] Schneider, W. (1985). Training high performance skills: Fallacies and guidelines. *Human Factors*, 25, 285-300.

FURTHER READING

- Hoffman, R.G., & Morrison, J.E. (1987). *Requirements for a device-based training and testing program for M1 gunnery. Volume 1: Rationale and summary of results* (Report No. FR-TRD-87-41). Alexandria, VA: Human Resources Research Organization.
- Lane, N.E. (1987). *Skill acquisition rates and patterns: Issues and training implications*. New York: Springer-Verlag.
- Lee, T.D., & Magill, R.A. (1983). The locus of contextual interference in motor skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 730-746.
- Fraisse, P. (1984). Perception and estimation of time. In M.R. Rosenzweig & L.W. Porter (Eds). *Annual Review of Psychology*, 35, 1-36.
- Shea, J.F., & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 179-187.
- Shiffrin, R.M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127-190.
- Stevens, S.S. (1975). *Psychophysics: Method, theory, and application*. New York: Wiley.

ACKNOWLEDGMENTS

This research was supported by a grant from the NASA Dryden Flight Research Center (DFRC). Special thanks to Jack Kolf for his kind assistance in tracing the history of "fast time" simulation back to Jack Kolf at NASA Dryden. Kolf was the contract monitor for this effort, and he was the first to originate the concept of "fast time" simulation while working on the NASA X-15 project. His encouragement and support of our research efforts began before the actual contract. His prior experience and suggestions were invaluable. NASA Dryden's chief pilot, Rogers Smith experimental suggestions greatly aided this research.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
1. AGENCY USE ONLY (leave blank)	2. REPORT DATE February 1994	3. REPORT TYPE AND DATES COVERED Conference Publication	
4. TITLE AND SUBTITLE Technology 2003, Volume 2		5. FUNDING NUMBERS	
6. AUTHOR(S) Michael Hackett, Compiler			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Technology Transfer Program, Code CU		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546		10. SPONSORING / MONITORING AGENCY REPORT NUMBER NASA CP-3249, Vol. 2	
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Unclassified/unlimited Subject Category 99		12b. DISTRIBUTION CODE	
13. ABSTRACT Proceedings from symposia of the Technology 2003 Conference and Exposition, December 7-9, 1993, Anaheim, CA. Volume 2 features papers on artificial intelligence, CAD&E, computer hardware, computer software, information management, photonics, robotics, test and measurement, video and imaging, and virtual reality/simulation.			
14. SUBJECT TERMS Artificial intelligence, computer aided design, computers, computer programs, photonics, robotics, video equipment, imaging techniques, computerized simulation, virtual reality		15. NUMBER OF PAGES 483	
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited

Available from NASA Center for AeroSpace Information
800 Elkridge Landing Road
Linthicum Heights, MD 21090-2934
(301) 621-0390

